Progetti Intelligenza Artificiale 2 (IA2) a.a. 2017-2018

# Premesse

I *Big Data* hanno caratteristiche di volume, velocità, varietà e veridicità che spesso eludono le tecniche e le architetture di elaborazione tradizionali. C’è, quindi, un grande interesse nello sviluppo di nuove architetture in grado generare valore dai Big Data sfruttandone le caratteristiche peculiari.

Il progetto IDRA (Inductive Deductive Reasoning Architecture) [[1](#Ene17)] si pone nel solco di questa linea di ricerca, proponendo una *famiglia di architetture* per i Big Data che combinano ragionamento deduttivo (tipico delle ontologie) e ragionamento induttivo (tipico dei metodi empirici), in modo da beneficiare delle potenzialità dei due tipi di ragionamento e compensarne le rispettive deficienze.

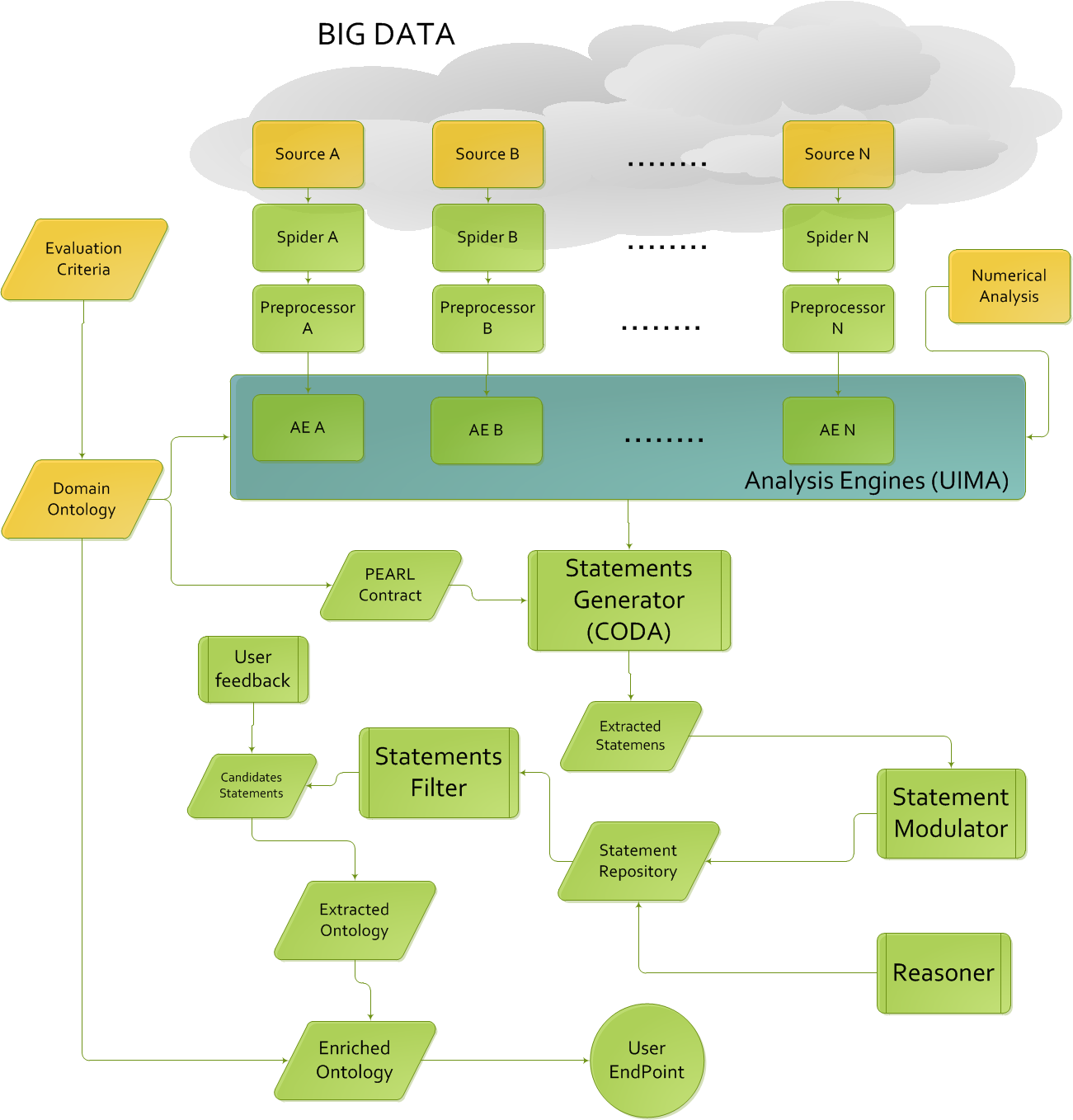


Figura 1 IDRA

La scelta di uno specifico membro (o *istanza*) della famiglia IDRA (vedi Figura 1) per uno specifico dominio ed applicazione si effettua mediante la scelta di opportune *fonti* di informazione (ciascuna associata ad uno spider, un preprocessore, ed un analizzatore linguistico/semantico), un’*ontologia di dominio,* un insieme di *criteri di valutazione* (che esprimono come tradurre l’informazione estratta in nuovi fatti nella base di conoscenza) ed un’*analisi numerica* (che esprime l’attendibilità delle fonti e dei metodi di analisi impiegati).

L’elaborazione delle fonti di informazione produce delle annotazioni (con relativi livelli di confidenza) che sono utilizzati dal *generatore di statement* per produrre nuove triple RDF. Internamente il generatore si avvale del sistema CODA [[2](#Fio14)] e del relativo linguaggio di trasformazione PEARL [[3](#Paz121)], che permettono di disaccoppiare lo schema delle annotazioni e l’ontologia di destinazione.

La confidenza delle triple generate viene inizialmente stabilita sulla base della confidenza delle fonti e dei metodi impiegati, mentre viene successivamente *modulata* sulla base di nuove evidenze raccolte dal sistema. IDRA, infatti, assume che le fonti di informazioni siano strutturate come flussi in continuo divenire, tali per cui il sistema deve rimanere pronto e reagire ad ogni nuovo elemento informativo appena esso diventa disponibile.

Un *filtro* sulle triple accumulate permette di selezionarne un sottoinsieme che al momento dell’interrogazione costituisce un’*aggiunta* plausibile alla base di conoscenza sottostante.

# *Topic 1*: Annotazione di sorgenti di informazione (1-2 persone)

Molti siti di informazione sono ormai dotati di *feed* RSS o Atom per permettere agli utenti di rimanere aggiornati sulle ultime notizie pubblicate (spesso in un settore a scelta dell’utente, per esempio quello delle notizie riguardanti la medicina e la salute).

L’obiettivo del progetto è realizzare un componente per la gestione di feed, che reagisca agli aggiornamenti, attivando il processo di analisi.

Il componente realizzato deve almeno:

* permettere di indicare un feed;
* permettere di indicare la frequenza di aggiornamento del feed;
* permettere di indicare un preprocessore (come annotatore UIMA);
* permettere di indicare un analizzatore linguistico/semantico (come annotatore UIMA);
* permettere di indicare un file PEARL contenente le regole di trasformazione per la generazione di triple a partire dalle annotazioni prodotte per mezzo dall’analizzatore linguistico/semantico;
* *opzionalmente:* utilizzare la conoscenza acquisita per una qualche applicazione.

A titolo di esempio, si può pensare di riconoscere menzioni di risorse definite in DBpedia (<http://dbpedia.org/>) servendosi del sistema DBpedia Spotlight (<http://www.dbpedia-spotlight.org/)>, e di memorizzare tali menzioni all’interno di un triplestore usando il vocabolario Web Annotation (<https://www.w3.org/TR/annotation-vocab/>). Riguardo alla parte opzionale, si può supportare l’analisi delle notizie (es. frequenza delle diverse menzioni, similarità tra notizie) e la ricerca semantica delle notizie (es. recuperare le notizie che hanno menzioni di istanze di determinate classi nella base di conoscenza). Un aspetto interessante di questo topic è il considerare la dimensione temporale, discutendo come essa possa influire ed essere sfruttata nei diversi task.

Ulteriori contributi possono essere proposti dal gruppo che sta svolgendo il progetto.

# *Topic 2*: Realizzazione di un Annotatore UIMA per articoli medici (2-3 persone)

Questo topic è simile al precedente, ma si differenza per l’input. In questo caso non si dovrà analizzare un insieme di feed RSS, ma gli studenti dovranno concentrarsi sugli articoli medici di una data patologia (ad esempio l’autismo). Per quanto riguarda i testi da analizzare, si potrà usare il portale PubMed (<https://www.ncbi.nlm.nih.gov/pubmed>), in quanto permette di avere accesso agli abstract dei vari paper. Altrimenti si potrebbe usare <https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/> per scaricare gli articoli nella loro interezza. Per l’autismo esiste invece il sito <https://ndar.nih.gov/> . BioPortal (<https://bioportal.bioontology.org/>) è un popolare repository di ontologie biomediche, che offre un servizio di annotazione automatica di testi (<https://bioportal.bioontology.org/annotator>). Utilizzando questo servizio, è possibile identificare le ontologie da utilizzare nel task.

Il componete sviluppato deve essere in grado di recuperare direttamente il contenuto da analizzare sulla base di un criterio di ricerca indicato dall’utente. In aggiunta, il sistema dovrebbe verificare la presenza di aggiornamenti, sfruttando i meccanismi offerti dall’archivio utilizzato. Come minimo, il sistema dovrebbe essere in grado di rilanciare periodicamente la medesima ricerca, e verificare la presenza di nuovi risultati.

In modo simile a quanto descritto nel primo topic, l’ontologia potrà essere usata per annotare menzioni di concetti negli abstract o nei testi, e le annotazioni potranno essere usate ai fini di produrre analisi, visualizzazioni e supportare la ricerca semantica. Un aspetto interessante di questo dominio è l’eventuale necessità di usare contemporaneamente più ontologie e ciò che esso comporta (es. necessità di allineare le ontologie).

# *Topic 3*: Raccolta, analisi ed utilizzo di metadati su Linked Dataset (2-3 persone)

Il paradigma del Linked Data ha promosso la pubblicazione e l'integrazione di dati attraverso l'uso di un modello di dati a grafo in cui è possibile creare collegamenti tra risorse, anche appartenenti a dataset diversi. Ciò è analogo a quanto accade nel Web dei Documenti dove i link ipertestuali possono puntare ad una pagina all'interno dello stesso sito oppure a pagine esterne.

In aggiunta alla descrizione delle risorse del dominio, è possibile pubblicare descrizioni di ciascun dataset nel suo complesso. Queste descrizioni conformi al vocabolario VoID (<https://www.w3.org/TR/void/>) contengono metadati utili a capire come accedere al dataset (es. presenza di uno SPARQL endpoint), qual è la struttura del dataset (es. numero di risorse, triple, vocabolari) e come esso è collegato ad altri dataset. Usando questi metadati è possibile stabilire (anche automaticamente) se un dataset è rilevante per un certo scopo (es. copre il dominio di interesse) e, successivamente, come usare il dataset in questione.

Lo scopo del progetto è realizzare un sistema che permetta di gestire un catalogo di descrittori VoID, supportando la ricerca di dataset sulla base di diversi criteri e la generazione/visualizzazione di statistiche.

La costruzione del catalogo può essere effettuata mediante approcci diversi:

* caricamento di specifici VoID file
* partendo da una risorsa o da un dataset, ottenere il relativo VoID (sfruttando i meccanismi descritti qui: <https://www.w3.org/TR/void/#deploying>)
* partendo dai descrittori VoID caricati in precedenza, considerare in modo ricorsivo i VoID dei dataset che sono raggiungibili per mezzo di link

Si chiede, inoltre, di analizzare servizi di indicizzazione esistenti per stabilire se sia possibile usarli per recuperare metadati. Alcuni servizi di interesse sono:

* Data Hub: <https://datahub.io/>
* Linked Open Vocabularies (LOV): <http://lov.okfn.org/>
* voiD store: <http://void.rkbexplorer.com/>
* LOD Cloud: <http://lod-cloud.net/>
* LOD Stats: <http://stats.lod2.eu/>
* BioPortal: <https://bioportal.bioontology.org/>

Una parte importante del progetto è l’analisi dei metadati recuperati evidenziando aspetti quali:

* Tipologia dei metadati forniti nella descrizione dei dataset;
* Distribuzione dell’uso dei diversi tipi di metadati nella descrizione dei dataset;
* Distribuzione dei metadati.

Nello studiare i tipi di metadati forniti, si dovrà ragionare su come ciò possa influire sull’utilità dei metadati stessi nello svolgimento delle attività di scoperta e selezione dei dataset.

La ricerca di dataset rilevanti può, infatti, utilizzare diversi tipi di metadati e diverse tecniche, per esempio:

* Ricerca testuale all’interno delle descrizioni testuali dei dataset;
* Matching sulle keyword (opzionalmente supportata da risorse linguistiche);
* Uso dei topic indicati nelle descrizioni (opzionalmente supportata da tecniche di matching quando i topic sono ottenuti da schemi di classificazione diversi)

È possibile considerare l’idea di indicizzare il contenuto stesso di ciascun dataset, in modo da permettere di ricercare dataset contenenti determinate risorse di interesse.

# *Topic 4*: Acquisizione di informazione da file tabulari (1-2 persone)

Fogli di calcolo e file CSV sono spesso scelti da organizzazioni ed istituzioni per rendere accessibili i dati in loro possesso. Questa scelta è motivata dalla familiarità con i programmi per editare questo genere di file (es. Excel o persino un semplice editor di testi) e dal fatto che praticamente ogni sistema per la gestione di dati supporta l’esportazione in formato tabulare.

La convenienza e la duttilità dei formati tabulari è controbilanciata dal fatto che in generale non possono essere interpretati in maniera precisa, a meno di non aver ricevuto informazioni dettagliate sull’organizzazione dei (relazioni tra i) dati (non sempre espressi come una sequenza di record) e delle etichette (non solo delle colonne, ma in alcuni scenari anche delle righe).

Per incrementare l’interoperabilità (semantica) di questi dati tabulari può essere utile trasformarne il contenuto informativo in una base di conoscenza conforme agli standard del Semantic Web.

Lo scopo del progetto è sviluppare un componente per l’acquisizione di informazione da file tabulari.

Il componente realizzato deve:

* Supportare diversi formati (quantomeno deve essere data la possibilità a livello architetturale di aggiungerne il supporto);
* Supportare l’interazione con l’uomo al fine di guidare il processo di acquisizione;
* Sfruttare l’ontologia e/o informazione linguistica per facilitare l’interazione con l’uomo.

Lo sviluppo del progetto dovrà essere preceduto da un’attività di analisi del problema finalizzata all’identificazione e alla classificazione di diversi scenari. Ogni scenario sarà caratterizzato dalla presenza di diversi elementi di complessità, che possono essere determinati dai dati tabulari in ingresso, dall’ontologia che si intende usare nel processo di acquisizione e dal loro usato combinato.

Per quanto riguarda le ontologie, è possibile partire da quelle a carattere biomedico citate in [[4](#Ene171)]: AEO (Anatomical Entity Ontology: <http://obofoundry.org/ontology/aeo.html>) and DDANAT (Dictyostelium Discoideum Anatomy: <http://obofoundry.org/ontology/ddanat.html>). È possibile, inoltre, considerare altre ontologie recuperate tramite i servizi citati nei topic 2 e 3.

Per quanto riguarda il processo di mappatura in RDF, il componente può usare tecniche di elaborazione del linguaggio naturale e risorse linguistiche (es. WordNet) per migliorare la capacità del sistema di associare proprietà e classi alle colonne della tabella. Similmente, l’ontologia target può essere sfruttata per generare suggerimenti basati sulla compatibilità dei tipi in gioco con il dominio/range delle proprietà.

Gli studenti potranno trarre ispirazione dal sistema Sheet2RDF [[5](#Sheet2RDF15)] per la triplificazione di dati tabulari, che sfrutta l’architettura di acquisizione di conoscenza CODA ed il suo linguaggio di trasformazione PEARL. Sheet2RDF propone un approccio stratificato: nonostante alla base vi sia un potente linguaggio di trasformazione, in molti casi le regole di trasformazione vengono generate automaticamente dal sistema mediante euristiche attivate dal riconoscimento di pattern nei file di input. In questo processo l’utente può intervenire fornendo le informazioni mancanti per l’applicazione delle euristiche mediante un’interfaccia utente guidata dall’ontologia target, e solo in rari casi agendo direttamente sulle regole di trasformazione. Sheet2RDF verrà integrato nelle prossime release dell’editor collaborativo di tesauri ed ontologie Vocbench 3 [[6](#Ste17)].

# *Topic 5*: Estrazione di porzioni di ontologie di grandi dimensioni (2-3 persone)

Molto spesso la dimensione di una base di conoscenza è determinata dalle asserzioni di fatti particolari riguardanti un grande numero di istanze appartenenti all’universo del discorso. In altre circostanze, l’elemento determinante è la parte concettuale della base di conoscenza che descrive in modo ricco un grande numero di concetti e di relazioni tra di essi.

Basi di conoscenza di grandi dimensioni sono spesso difficili da esplorare, interrogare e, più in generale, elaborare. Quest’ultima difficoltà è connessa non solo alla complessità computazionale degli algoritmi usati ma anche alla difficoltà dell’intervento umano: per esempio, allineare due ontologie di grandi dimensioni è una task computazionalmente difficile, ma non meno difficile è la valutazione del risultato da parte di esperti umani, specialmente se si vuole un allineamento completo (che teoricamente richiede di esplorare il prodotto cartesiano delle due ontologie).

In molti casi, tuttavia, non c’è realmente bisogno di usare una base di conoscenza nel suo complesso, perché ne è sufficiente soltanto una porzione.

Lo scopo del progetto è progettare e realizzare un sistema per l’estrazione di porzioni di ontologie sulla base di criteri di inclusione riguardanti classi, proprietà ed individui. Al gruppo che intende lavorare su questo topic è chiesto di analizzare il problema e proporre algoritmi, possibilmente informati dalla letteratura scientifica corrente, valutandone vantaggi e svantaggi. Un’altra parte importante del progetto è l’interfaccia utente attraverso cui selezionare la porzione di ontologia di interesse. A titolo di esempio, l’utente dovrebbe essere messo in grado di effettuare delle ricerche all’interno dell’ontologia, per selezionare le entità di interesse, costruendo in modo incrementale la porzione di ontologia da estrarre.

# *Topic 6*: Attività trasversali condivise nei vari topic (1-2 persone)

Lo scopo è quello di individuare un tema che sia di utilità in più di un topic e quindi fornire una attività di supporto in ciascuno di essi (es. GUI, Gestione delle ontologie, triplificazione, elaborazione linguistica, etc.) In tal modo si permetterebbe di aumentare la sinergia dei flussi di processing semantico dei Big Data all’interno della stessa architettura.

# Requisiti Comuni

I progetti possono essere svolti anche in gruppi a numerosità da definirsi di volta in volta.

Il risultato dell’attività progettuale dovrà essere discusso durante la prova di esame orale, avvalendosi del supporto di una presentazione (di circa 20 minuti).

La presentazione dovrà essere consegnata (inviata per posta elettronica) con qualche giorno di anticipo.

I sorgenti del progetto accompagnati da una breve relazione tecnica (circa 10 pagine) dovranno essere consegnati almeno una settimana prima della prova di esame. Durante questo lasso di tempo, il gruppo potrà ricevere richieste di chiarimento o di modifiche al progetto nel caso di gravi lacune che pregiudichino l’ammissione alla prova orale.

La presentazione e la relazione tecnica possono essere scritti in italiano o in inglese a scelta e non devono rispettare un formato prestabilito.

La data d’esame potrà essere concordata direttamente da ciascun gruppo ed essere diversa da quella degli altri. In ogni caso tutte le presentazioni richiedono la presenza di tutti gli studenti (sia presentanti il risultato del proprio lavoro che solo uditori).

# Bibliografia

x

|  |  |
| --- | --- |
| 1. | Enea, R.: IDRA: An ontology driven Cognitive Computing System. ISWC 2017 Doctoral Consortium (2017) |
| 2. | Fiorelli, M., Pazienza, M.T., Stellato, A., Turbati, A.: CODA: Computer-aided ontology development architecture. IBM Journal of Research and Development 58(2/3), 14:1,14:12 (March-May 2014) |
| 3. | Pazienza, M.T., Stellato, A., Turbati, A.: PEARL: ProjEction of Annotations Rule Language, a Language for Projecting (UIMA) Annotations over RDF Knowledge Bases. In : LREC, Istanbul (2012) |
| 4. | Enea, R., Pazienza, M.T., Turbati, A., Vitiello, D.: Conceptual knowledge management: the case of biomedical information. In : Proceedings of the third edition of the Joint Ontology Workshops (JOWO-17). September 22-23, 2017, Bolzano, Italy (2017) (in press). |
| 5. | Fiorelli, M., Lorenzetti, T., Pazienza, M.T., Stellato, A., Turbati, A.: Sheet2RDF: a Flexible and Dynamic Spreadsheet Import&Lifting Framework for RDF. In Ali, M., Kwon, Y.S., Lee, C.-H., Kim, J., Kim, Y., eds. : Current Approaches in Applied Artificial Intelligence (Lecture Notes in Computer Science) 9101. Springer, Cham (2015), pp.131-140 |
| 6. | Stellato, A., Turbati, A., Fiorelli, M., Lorenzetti, T., Costetchi, E., Laaboudi, C., Van Gemert, W., Keizer, J.: Towards VocBench 3: Pushing Collaborative Development of Thesauri and Ontologies Further Beyond. In : 17th European Networked Knowledge Organization Systems (NKOS) Workshop (21st September 2017), Thessaloniki, Greece |

x