**Ontology learning**

**Cosa significa?**

Ontology learning rappresenta l’apprendimento in maniera automatica o semi automatica di una ontologia. Una ontologia computazionale è un formalismo di rappresentazione della conoscenza di un dominio di interesse strutturata in entità e relazioni tra entità. La conoscenza espressa è di tipo gerarchico.

Si può pensare all’ontology learning come il processo inverso della creazione manuale di un’ontologia, ovvero un processo di reverse engeenearing. Le persone definiscono un mondo che hanno in testa e lo traducono e lo utilizzano per creare documentazione e per distillare conoscenza scritta dalla quale si può fare questo processo inverso per tornare alle concettualizzazioni originali. Come si recupera il pensiero originario.

**2 problemi:**

* La Conoscenza del mondo non è totalmente codificata. non tutta la conoscenza del mondo è scritta nelle enciclopedie, nelle basi documentali di partenza. Noi conosciamo cose implicite che provengono dal senso comune, dall’esperienza o da altro che non si ritrovano codificate in nessuna risorsa e non si può neanche pensare di codificare tutto
* La conoscenza del dominio che si deve codificare non è utilizzata completamente. La domain knowledge potrebbe non essere usata completamente
* Si crea un divario e quello che si fa è cercare il compromesso tra usabilità e task considerati.

Differenze con altre metodologie:

* **Ontology learning**: creazione automatica di una ontologia a patire dai testi e dai documenti a disposizione
* **Ontology population**: fa parte dell’ontology learning ma si occupa di popolare una ontologia già creata e cerco di prendere dei concetti da una base documentale e inserirli all’interno di concetti e delle relazioni gia esistenti. Una sorta di cluster. Ad esempio ho il concetto di mela, trovo dei tipi di mele e cerco di inserirli nell’ontologia
* **Ontology based Annotation:** L’attenzione non è sull’ontologia, ma sui documenti. Voglio creare una base documentale arricchita evidenziando nel testo termini o pezzi di frasi che riflettono dei concetti di una ontologia di riferimento. Quindi c’è sempre un riferimento ad una ontologia già esistente
* **Ontology enrichment:** cerco di arricchire l’ontologia a partire da una base documentale. Lo scopo è cercare di capire cosa manca in quell’ontologia, non popolarla. Ad esempio nuovi concetti e nuove relazioni. Automaticamente si cerca di capire sottoalberi di una ontologia mancanti. Si modifica un’ontologia gia esistente.

Categorizzazione dei vari livelli di profondità nella definizione di ontologie: Ontology learning stack. Passi per la creazione di un’ontologia

* Base documentale
* Creazione di un dizionario di termini
* Creazione di un glossario. Rispetto al dizionario è un elenco di termini con glossa
* Tesaurus: simile a wordnet. Esistono relazioni orizzontali di sinonimia
* Tassonomie
* Ontologia
* Rappresentazione logica

Categorizzazione dei task relativi all’ontology learning:

* **Term Extraction:** mi fermo alla terminologia. Trovare nomi di concetti e di relazioni. Come ad esempio termini ricorrenti. Una tag cloud è una term extraction visualizzata.
* **Synonym extraction:** andiamo in direzione di teaurus per estrarre i sinonimi di termini
* **Concept extraction:** si fa un gloss learing, elenco di parole che identificano un argomento. Si vanno ad estrarre concetti nelle loro intenzioni ed estensioni, cioè dato un concetto andare a trovare la definizione di quel concetto (intenzione o gloss learning) e per estensioni, tutta la terminologia utilizzabile nei contesti per riferirsi ad un determinato concetto.
* **Concept Hierarchies extraction:** induco direttamente gerarchie di concetti, quindi si estraggono relazioni iperonimiche.
* **Relation extraction:** estraggo relazioni semantiche generiche
* **Population:** Information extraction, quindi da un concetto si può fare information extraction, o robe banali come espressioni regolari per estrarre codice fiscale. Pattern learning, NER

Metodi principali per fare ontology learning:

* Natural language processing **NLP**
* Formal concept analisys **FCA**
* Machine learning **ML**

Quando si parla di NLP, ci sono diversi metodi. Il più semplice di tutti è l’utilizzo di Part of speech. L’uso di POS identifica dei possibili concetti e relazioni.

* I verbi tipicamente sono relazioni tra concetti
* I sostantivi sono tipicamente concetti
* I nomi propri sono individui
* Gli aggettivi sono attributi di concetti

Posso usare il preprocessing classico di una base documentale come POS tagging, stemming, lemmatizzazione, normalizzazione, denormalizzazione come in NER. Posso aggiungere un’analisi sintattica, quindi uso le dipendenze sintattiche, posso usare similarità su vettori, tutto ciò che abbiamo visto nella semantica distribuzionale, quindi similarità TF-IDF. Posso usare risorse linguistiche come WordNET, BabelNET, ConceptNET...

**Formal Concept Analisys (FCA)**

L’FCA è un metodo che proviene dall’ambito matematico, simbolico, molto usato per fare ontology learning e soprattutto per fare estrazione automatica di tassonomie. Prima di introdurre FCA, bisogna dire che cosa sono:

* **Contesto formale**
* **Concetti**
* **Features**
* **Matrice di adiacenza**



Usiamo l’esempio:

* **Figura A:** La prima figura è una matrice organizzata per concetti e features. I concetti sono sulle colonne, le features sulle righe. Si può pensare di rappresentare un concetto attraverso una serie di features. E’ una cosa già vista, nel Text mining, o nella semantica distribuzionale. Codifico i concetti attraverso una serie di features. Queste features sono booleane dicono se un concetto ha o meno una features.
* **Figura B:** La figura B è **un contesto formale**. Che cosa è un contesto formale? Si ha una lista di concetti (C1, ..... C7) e in più si aggiungono 2 simboli di testa e coda. Se prendiamo il concetto1 vengono associati 2 insiemi: {f1,f2} e {c4, c7}. Sono basati sulla tabella A. Che cosa sono? {f1 e f2}? Se prendiamo C7, notiamo che ha tutte le features. Rappresenta il concetto che ha tutte le features quindi è posto sulla testa del lattice.
* Figura C: Lattice: é una sorta di grafo. Parte dal concetto C7, quello che ha tutte le 3 features, poi scendo al secondo livello e trovo tutti i concetti che hanno solo 2 features, poi scendo ulteriormente e ottengo i concetti con una sola feature. Infine ho l’elemento di coda, in cui nessun concetto è presente poichè tutti i concetti hanno almeno una feature. Il lattice ha questa forma in cui si espande per tutte le combinazioni possibili poi pian piano si riduce di nuovo perchè il livello penultimo contiene ogni singola features. Se ho 10 features, il penultimo livello avrà 10 elementi.

Che cosa fa FCA?. E’ un metodo formale e deterministico che prende la matrice A e costruisce il formal context da cui si può ricavare il lattice. Quei concetti C1, C2,C3....C7 sono concetti latenti, ovvero che sono stati estratti autonomicamente che hanno in se un aspetto:

* Intensionale: sono le features che lo contraddistinguono
* Estensionale: sono gli elementi della base di dati iniziale che rientrano nel sotto spazio di features

Cosa serve? Il formal concept può essere vista come una tassonomia, perchè da una parte abbiamo tutte le features che vengono condivise in maniera sequenziale, ad esempio Frutto, mela e pera: Frutto è la superclasse, mela e pera sono sotto classi che ereditano delle proprietà e ne hanno altre loro, pera e mela hanno forme diverse e da li si sottospecificano. Si può usare FCA come meccanismo preciso per creare tassonomie. **FCA meccanismo per indurre automaticamente delle tassonomie.**

Machine Learning:

tutti meccanismi supervisionati per fare ontology learning come:

* Classificatori bayesiani
* Decision trees
* Support vector machines
* Reti neurali

Open Information extraction:

Si tratta di una specie di ontology learning. L’ontology learning è anche definito come il passaggio da un dato non strutturato come il testo a un dato strutturato in maniera automatica. L’open information extraction viene anche chiamato come shallow ontology learning perche crea informazione semi-strutturata. Questa semi struttura è ad esempio quella della teoria di patrick hanks: prendere delle frasi e tira fuori triplette, che nella maggio parte sono **arg1 - verbal phrase - arg2**

Indica un approccio che parte dall’information extraction e quell’Open sta ad indicare che un approccio che è quello di estrarre tanto da tantissimi dati. Quindi è nato per rendere scalabili, veloci, e approssimativi i sistemi di Information extraction. Se ho un meccanismo che struttura un minimo il testo riesco a interrogare basi documentali ampie per capire chi ha fatto cosa. Quindi costruire rappresentazioni semi strutturate.

La maggior parte dei sistemi usa:

* POS, dependency parsing
* Word sense disambiguation
* Regole su cosa possono essere gli argomenti

**Problemi di OIE:**

Ognuno di noi può inventare un sistema che estragga triple, basta che uso POS, uso dei vettori con dei filtri... ho molti tipi di estrazioni ed è difficile valutare l’OIE. Non c’è un meccanismo unanime per valutare e comparare i sistemi di OIE

**PROGETTO:**

Link libreria python per fare **FCA** ovvero prendere una matrice di adiacenza: concetti-features e creare **il formal context**, eventualmente da visualizzare. Su cosa applicare l’FCA?

* Informazioni testuali
* Estrarre termini da una lingua e come features possiamo usare delle dipendenze sintattiche con altre parole, ad esempio ho un intorno lessicale oppure delle proprietà che estraiamo da altre risorse come propery norms. Ho un dataset e lo riscrivo come una matrice di adiacenza dove metto i concetti del dataset e nelle features tutte le property norms nel dataset. Quando coccodrillo ha “is dangerous” mettiamo X. Una volta costruita la matrice, la si può importare con un csv. Uno degli algoritmi della libreria estrae il formal concept.