Università Ca' Foscari di Venezia

Linguistica Informatica Mod. 1

Anno Accademico 2010 - 2011

Informazione e conoscenza

Rocco Tripodi

rocco@unive.it

Informazioni testuali

Internet ha fatto crescere il bisogno di applicazioni per il trattamento automatico del testo

I contenuti sono veicolati attraverso il linguaggio naturale

Le pagine web contengono per lo più informazioni non strutturate

plain text VS array, tabelle, alberi e reti

In informatica l'informazione strutturata si ottiene dall'analisi dei dati e dalla loro suddivisione in componenti inserite in categorie

Informazioni che per quanto riguarda i dati testuali, possono essere ricavate tramite apposite metodologie di NLP o mediante l'elaborazione di dati strutturati come per esempio tabelle e infobox

Classificare i documenti

Document retrieval

Il WWW consiste prevalentemente di documenti che devono essere cercati e classificati per essere resi disponibili agli utenti che li cercano
Architettura della rete internet e *web crawlers*

Text categorization

Processo che suddivide i documenti in base al loro contenuto, invece il

Text classification

È un processo più generale che suddivide i documenti in base a delle proprietà generali come l'autore, la lingua, ecc.

Un documento può appartenere a più classi

Le classi possono essere organizzate in una struttura genarchica

Gli scopi delle classificazioni

Routing

Una fonte di informazioni vuole suddividere i messaggi creati per poterli smistare ai clienti

Indexing

Le biblioteche digitali devono classificare i loro contenuti in base ad un vocabolario controllato, per agevolarne la fruizione. Indicizzazione manuali sono molto onerose da adottare per grandi moli di dati.

Sorting

Per fare ordine in una collezione di documenti indiscriminata

Come classificare 1

Ci sono diversi fattori che caratterizzano i dati

Granularità
: Quante categorie bisogna utilizzare? Si parla di filtraggio a grana grossa quando si usano poche categorie, per cui vengono creati dei *cluster* abbastanza corposi.

Dimensionalità
: Quante caratteristiche bisogna considerare? Se ogni parola del documento rappresenta una caratteristica allora si opererà su una dimensionalità elevata che sarà popolata in modo sparso dai documenti. Al contrario se si decide di utilizzare un vocabolario controllato la dimensionalità si ridurrà in funziona della dimensione del vocabolario.

Come classificare 2

Esclusivity

A quante categorie appartiene un documento? Il documento deve appartenere a classi disgiunte o per facilitarne la ricerca c'è bisogno che compaia in più nodi dell'albero gerarchico?

Topicality

il documento riguarda un solo argomento o diversi?

Altre attività incluse nel processo di classificazione sono:

Document management

Convertire i documenti in un formato comune e taggarli uniformemente.

Concept mamagement

Nel mondo dei *new media*, c'è bisogno che le classi concettuali vengano continuamente aggiornate per far fronte alla popolarità di un argomento

Taxonomy management

Le tassonomie organizzano i documenti per il *browsing* e il *searching*

Rappresentazione della conoscenza 1

Categorie

Tipologia generale in cui possono venire raggruppati i concetti da un punto di vista logico.

Teoria dei predicabili (Aristotele)

Modi in cui le categorie possono essere attribuite a un soggetto
le suddivisioni sono organizzate in maniera gerarchica e partono dai concetti universali fino ad arrivare ai concetti particolari

Albero porfiriano

Sussunzione: ricondurre un concetto particolare nell'ambito di uno generale
proprietà binarie da verificare su un determinato oggetto, al fine di decidere la giusta collocazione (e l'ornitorinco?)
Problema: l'attribuzione di un oggetto ad una determinata categoria non avviene per la condivisione di determinate caratteristiche, bensì sull'aderenza o meno dell'oggetto a un'immagine mentale che ciascuno di noi ha di tale classe.
Le distinzioni sono accidenti dell'oggetto e come tali possono essere infinite.

Yahoo! Directory come esempio

Arts & Humanities
Photography, History, Literature

Business & Economy
B2B, Finance, Shopping, Jobs

Computers & Internet
Hardware, Software, Web, Games

Education
Colleges, K-12, Distance Learning

Entertainment
Movies, TV Shows, Music, Humor

Government
Elections, Military, Law, Taxes

Health
Diseases, Drugs, Fitness, Nutrition

News & Media
Newspapers, Radio, Weather, Blogs

Recreation & Sports
Sports, Travel, Autos, Outdoors

Reference
Phone Numbers, Dictionaries, Quotes

Regional
Countries, Regions, U.S. States

Science
Animals, Astronomy, Earth Science

Social Science
Languages, Archaeology, Psychology

Society & Culture
Sexuality, Religion, Food & Drink

Rappresentazione della conoscenza 2

Enciclopedia (Eco, U., Dall'albero al labirinto, Bompiani, Milano, 2007)

A differenza dell'albero che pretende di usare i termini delle sue disgiunzioni come disgiunzioni non ulteriormente definibili, il nodo enciclopedico rinvia a nozioni che lo definiscono e che saranno esposte nel corso della trattazione completa del concetto.

Con l'età moderna si passa dalle classificazioni della realtà alle classificazioni intorno al sapere che della realtà si ha. Le prime definizioni di enciclopedia e di sapere sono accompagnate da parole come *selva* e *labirinto*, per evidenziare la natura non ordinata dell'enciclopedia, all'opposto dell'albero gerarchico.

Labirinto *unicursale:* impone una direzione obbligata. Se fosse srotolato prenderebbe la forma di un filo

Labirinto manieristico o Irrweg che propone scelte alternative e tutti i percorsi portano ad un punto morto tranne una, se venisse srotolato prenderebbe la forma di un albero

DBpedia come esempio

DBpedia è la versione strutturata di Wikipedia

Tutte le informazioni presenti negli *infobox* di Wikipedia sono state estratte e trasformate in fatti RDF

Le informazioni presenti nel free-text non sono state trattate ma semplicemente riportate

Ad oggi la base di conoscenza di DBpedia descrive più di 3,4 milioni di entità, delle quali 1,5 milioni sono classificate in una ontologia consistente. Le entità includono 312,000 persone, 413,000 luoghi, 94,000 album musicali, 49,000 film, 15,000 video game, 140,000 organizzazioni, 146,000 specie e 4,600 malattie.

L'intento del progetto è quello di creare un punto di riferimento per il così detto Web dei dati; che consisterebbe creare dei riferimenti semantici per le entità presenti sul web in modo tale che possano essere identificate in maniera non ambigua.

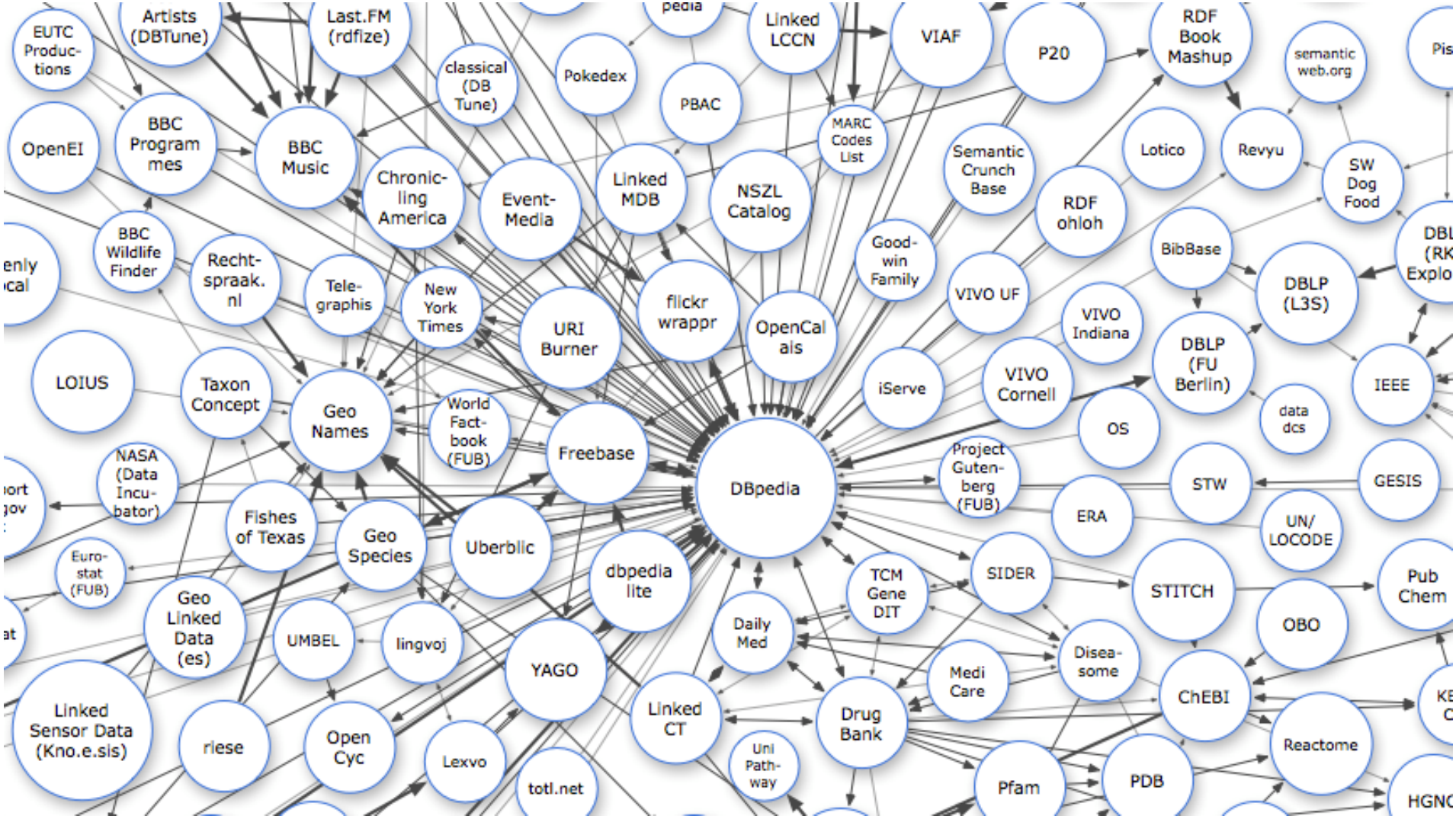
Rappresentazione della conoscenza 3

Rete

Ogni punto può essere connesso a qualsiasi altro, non può essere srotolato perché non ha né interno né esterno
Tale strumento è stato pensato per rappresentare l'organizzazione dei significati delle parole all'interno della mente (Quillian 1968). Questo approccio prevede che i concetti all'interno della mente umana siano organizzati in base ad una struttura reticolare.

Quillian introduce il concetto di rete semantica per definire gli elementi di una enciclopedia. Ogni nodo dell'enciclopedia può essere o un nodo tipo o un nodo occorrenza.
Un nodo tipo (*type*) può essere definito da una serie di altri nodi (*tokens*) che a loro volta in altri contesti possono diventare *type* per essere a loro volta definiti in un processo a catena.

LOD Cloud come esempio



Information Retrieval

Retrieval = indexing + search

L'*information retrieval* comincia con l'indicizzazione dei documenti. L'indice consiste in una lista di parole (normalizzate) contenute nella collezione di documenti

Ricerca mirata in funzione del contenuto (*querying*)

- classificazione argomentale
- strutturazione del contenuto
- indicizzazione

Ricerca esplorativa (*browsing*)

- navigazione
- approssimazioni successive
- storia e orientamento

Indicizzazione 1

Indice

L'indice di un libro consiste in una serie di parole (titolo) collegate al numero di pagina nella quale il loro contenuto viene trattato
Consente di esplorare il contenuto di un testo in base agli argomenti rilevanti

Documenti elettronici

Sono possibili ricerche sull'intero contenuto del documento. Viene ricercato l'esatto *matching* delle parole usate nella ricerca.

Stop words

Oltre alle indicizzazioni del *full text* possono essere effettuate indicizzazioni che escludono le parole che hanno un significato lessicale basso o ambiguo.
Comprendono elementi come gli articoli e le preposizioni

Un indice consiste nell'elenco di tutte le parole presenti nella collezione di testi.

Prima di essere inserite nell'indice le parole vengono lemmatizzate (entrata del vocabolario)

Indicizzazione 2

DocCount

Indica in quanti documenti il lemma occorre.

FreqCnt

Indica la frequenza del lemma

DocNo

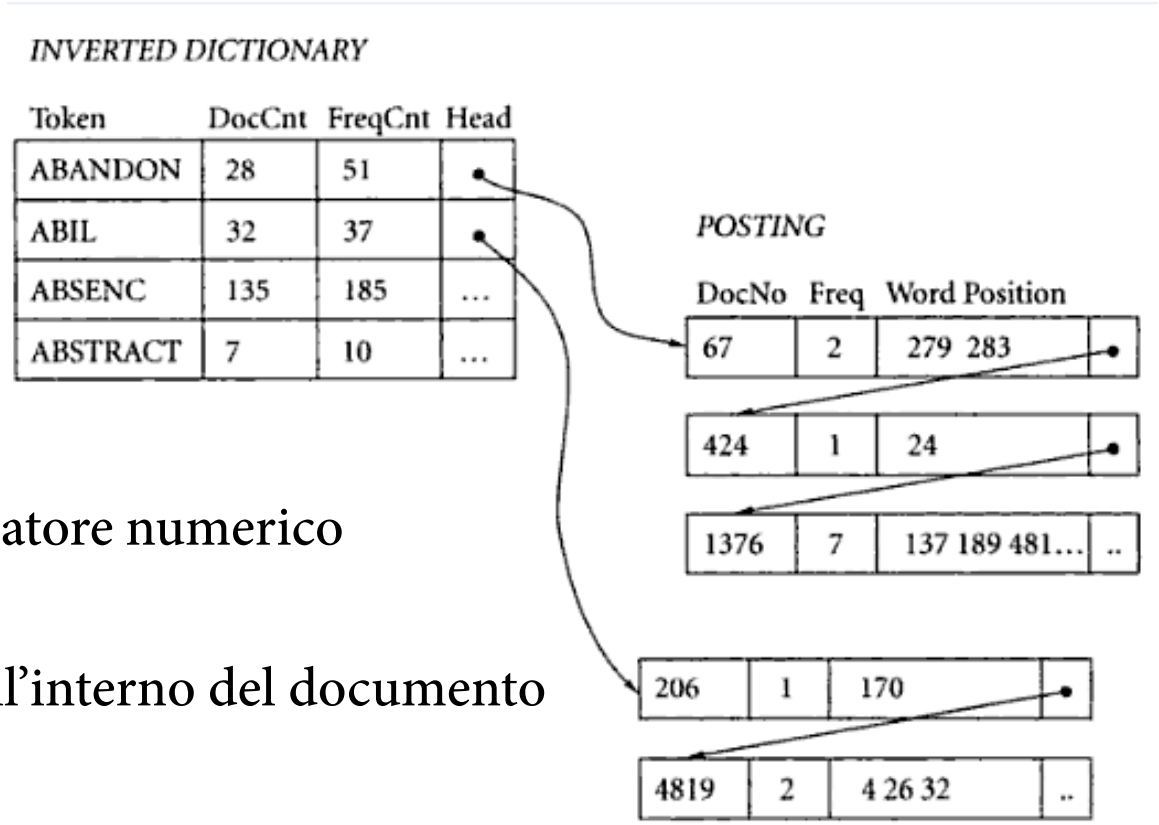
Ogni documento ha un identificatore numerico

Freq

Indica la frequenza del lemma all'interno del documento

Word Position

Indica la posizione del lemma all'interno del documento in termini di punti tipografici



Queries

Operatori booleani

I primi sistemi di indicizzazione oltre all'a ricerca dell'occorrenza ortografica consentivano di utilizzare gli operatori booleani per effettuare un sistema di filtraggio dei risultati

Table 2.1 Boolean truth tables

and	true	false
true	true	false
false	false	false

or	true	false
true	true	true
false	true	false

not	
true	false
false	true

Oltre agli operatori booleani i sistemi moderni devono espandere le *queries* utilizzando dei *thesauri* che comprendano sinonimi, iponimi, iperonimi e che riescano a fare i conti con la polisemia, l'ambiguità e le varianti.

Rilevanza delle queries

Secondo l'approccio che stiamo studiando il fatto che un documento contiene molte occorrenze di un determinato termine di ricerca, indica che quel documento è rilevante per quella particolare *query*.

Per calcolare la rilevanza di una *query* quindi si deve considerare da una parte la frequenza del termine all'interno del documento che stiamo cercando di classificare e dall'altra la frequenza dello stesso termine all'interno della collezione di documenti che abbiamo a disposizione

Inverse document frequency

$$Idf=\log(N/n)$$

N è il numero di documenti e n il numero di documenti in cui appare il termine

Recall: R = a/n

a è il numero di documenti rilevanti trovati e **n** è il numero di documenti rilevanti per la query

Precision: P = a/m.

a è il numero di documenti rilevanti trovati e **m** è il numero dei risultati restituiti

Information Extraction 1

Lo scopo non è trovare documenti, ma trovare informazioni all'interno dei documenti non strutturati

Regular expression (*regex*): creano un metodo per specificare un linguaggio. Costituiscono un set di stringhe su tre operazioni: adiacenza, ripetizione e alternanza.

{Mr.|Mrs.|Ms.|Dr.} {A|B|C|…|Z}. LASTNAME

dove LASTNAME indica ogni selezione da un elenco di nomi.

La *pipeline* comune per usare le *regex* è la seguente:

Tokeninnaz -> tagging -> Regex Matches -> Template Filler -> Template Merger

Information Extraction

Fastus (Link) è un sistema di IE creato per estrarre informazioni riguardanti eventi terroristici. L'estrazione delle informazioni avviene tramite il riempimento di *template* una volta individuate le espressioni regolari. Dalla frase

Terrorists attacked the local mayor's home in Bogota

si può voler estrarre il gruppo nominale con l'espressione regolare

GN = DET MOD NOUN

Che restituisce *local mayor*.

Bisogna notare che i *parser* sviluppati per l'IE operano su parti determinate di frase e non vengono usati per individuare tutte le strutture linguistiche. Spesso sono necessarie delle analisi grammaticali per gestire documenti complessi ed evitare errori di assegnamento.

IE avanzato

L'operazione vista in precedenza ci ha consentito di individuare all'interno della frase una *definite description*

Articolo determinativo + frase nominale

Oltre a voler individuare nel testo la presenza di un riferimento ad un *leader* politico, possiamo voler individuare con precisione a chi è indirizzato il riferimento. Le *definite descriptions* hanno sempre un riferimento ad una entità già menzionata nel testo.

In un sistema di IE avanzato il riferimento può essere recuperato individuando l'entità che appartiene alla classe *mayor* e se ci dovessero essere ambiguità si andrebbe ad effettuare una restrizione sui risultati includendo anche Bogotà

Tripla RDF

dbpedia:Bogotá dbpedia-owl:leaderName dbpedia:Samuel_Moreno_Rojas

Text mining

Data mining: cerca di trovare dei modelli, delle strutture ricorrenti nei dati.

Text mining non è riconducibile all'*information retrieval*, *extraction* o *classification*.

> Esso mira alla scoperta di nuove informazioni e trasformare le informazioni in conoscenza, generando nuova informazione.

Analisi → Dati strutturati → Eleborazione dati → Informazioni

Una delle applicazioni principali è la *semtiment analysis* che mira a individuare le intenzioni, gli stati d'animo, i possibili comportamenti o le aspettative del parlante. L'analisi di una frase prevede l'attribuzione di una polarità. Esistono diverse applicazioni che fanno *sentiment analiysis* attraverso le API e Twitter