Le fasi di un processo di Text Mining

Information Retrieval

↓

Information Extraction

↓

Information Mining

↓

Interpretazione

L'Information Extraction (IE)

- Con *IE* si intende l'insieme di operazioni volte ad estrarre da un *corpus* la porzione di informazione ritenuta rilevante per l'obiettivo conoscitivo e a riversarla in in una base di dati strutturata
- Le espressioni del linguaggio naturale sono ambigue e anche documenti con una *struttura*, come le liste di annunci di lavoro su *web*, sono difficili da intepretare automaticamente
- Ancora più complicato è il compito di combinare informazioni provenienti da più documenti e registrarli in un'unica base di dati coerente

Sistemi di IE

- Alcuni sistemi di IE estraggono l'informazione da testi scritti in linguaggio naturale, ad esempio le periodiche Message Understanding Conferences hanno esaminato sistemi per estrarre informazione dai documenti più vari, quali, ad esempio, i rapporti relativi ad incidenti aerei
- Altri sistemi estraggono informazione da documenti strutturati o semi strutturati, come ad esempio *Citeseer* [Lawrence *et al.*, 1999] che costruisce basi di dati di pubblicazioni accademiche (*titoli, parole-chiave, citazioni, …)*

Information Extraction: Distilling Structured Data from Unstructured Text (1)

by Andrew McCallum, University of Massachusetts

«Gran parte dell'informazione nel mondo è rinchiusa all'interno di testi non strutturati, espressi in linguaggio naturale. Le tecniche di IE possono essere utili per liberarla»

Nel 2001 il Dipartimento statunitense per il lavoro si propose di costruire un sito *web* che aiutasse le persone a trovare opportunità di formazione continua in istituti, università e organizzazioni, nell'intero paese. Il Dipartimento voleva che questo sito fornisse in risposta a richieste booleane informazioni relative alla localizzazione, durata, calendario, prerequisiti, insegnanti, aree tematiche e descrizione dei corsi. Inoltre, era interessato nel *mining* del suo nuovo *database* per classificare e studiare gli andamenti della formazione negli Stati Uniti

Il progetto si basava quindi sull'integrazione di dati da fonti differenti, così da mettere insieme in maqniera automatica un'informazione dettagliata e strutturata da dichiarazioni provenienti da migliaia di istituzioni individuali, ogni tre mesi

Information Extraction: … (2)

Il primo e maggiore problema era che gran parte dei dati non era disponibile neppure in un formato semi-strutturato. Sebbene alcune organizzazioni più grandi avessero dei *database* interni relativi ai propri corsi, praticamente nessuno aveva degli interfaccia pubblici ai propri *database*. Le uniche informazioni disponibili universalmente erano le pagine Web progetttate per un *browsing* umano

Sfortunatamente e prevedibilmente, ogni ente aveva organizzato differentemente i testi. Alcune pagine Web contenevano tabelle testuali bidimensionali; altri avevano paragrafi strutturati per descrivere ciascun corso offerto, altri ancora utilizzavano una prosa in inglese, e ogni paragrafo descriveva liberamente un corso.

Occorreva, quindi, estrarre informazione strutturata da testi in lingua inglese, talvolta sotto forma di tabelle, oppure in prosa in stile libero. Vediamo il contesto in cui si colloca questa impressionante sfida, prima di vederne la soluzione

Information Extraction: … (3)

In letteratura si trovano articoli che trattano il problema di IE per dati semi-strutturati - dati in tabelle XML o CSV (comma separated value), non normalizzati, in schemi diversi, spesso con problemi di duplicazione dei record. Ma la maggior parte dell'informazione del mondo è ancor meno strutturata – espressa in linguaggio naturale – ad esempio, pagine Web, rapporti aziendali, articoli di giornali, rapporti di ricerca, e-mail, blogs, documenti storici

I testi possono essere cercati e ordinati efficacemente dai motori di ricerca, ma le tecniche di *data mining* o i sistemi di supporto alle decisioni richiedono un'elaborazione più fine. L'informazione racchiusa nel linguaggio naturale deve essere prima trasformata in *database* strutturati e normalizzati

IE ha per obiettivo il processo di riempire i campi e i record di un database partendo dall'informazione non strutturata, o in un vago formato. Allora, IE può essere visto come una fase precedente il *data mining* che successivamente scoprirà strutture nella base di dati

Information Extraction: … (4)

IE si articola in 5 fasi principali:

• la ***Segmentazione*** trova l'inizio e la fine delle parti di testo che andranno nel *database*. Nell'esempio del Dipartimento del Lavoro statunitense sulla formazione continua, trovare la prima parola e l'ultima del titolo del corso.

• la ***Classificazione*** determina quale campo del *database* è la corretta destinazione per ogni segmento di testo. Ad esempio, "Introduzione al CAD" appartiene al campo TITOLO DEL CORSO, mentre "Dr. Dallan Quass" nel campo dell'INSEGNANTE, e "Questo corso si rivolge..." nel campo DESCRIZIONE DEL CORSO. Spesso la- segmentazione e la classificazione sono eseguite insieme

• l'***Associazione*** determina campi appartengono allo stesso record. Ad esempio, alcuni corsi possono essere descritti da più paragrafi, e altri corsi soltanto da uno. Occorre, quindi, determinare quali paragrafi si riferiscono allo stesso corso. Questa fase talvolta viene chiamata come ***estrazione di relazioni***, perché si associano due entità

Information Extraction: … (5)

- ***La*** *Normalizzazione* mette l'informazione in un formato standardizzato in cui può essere facilmente confrontato.

Ad esempio, l'orario della lezzione può essere "2-3pm", oppure "3pm-4:30pm", o ancora "15.00-16.30" e si vogliono evitare sovrapposizioni. Ovviamente, il semplice confronto fra stringhe non funzionerebbe: i dati vanno convertiti in una rappresentazione standard (possibilmente numerica)

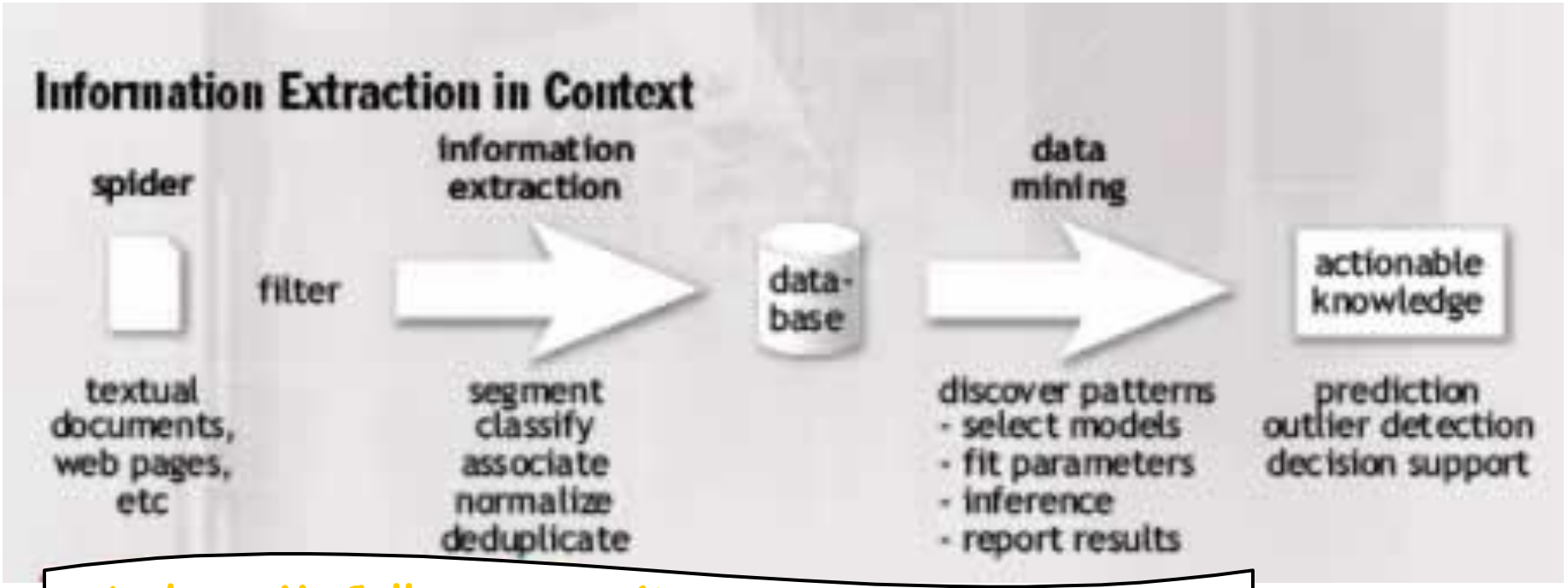
La *normalizzazione* è importante anche per le stringhe, ad esempio Giovanni Rossi e Rossi Giovanni renderebbe necessario definire un ordine fra nome e cognome

La *normalizzazione* è spesso collegata a punto successivo

Information Extraction: … (6)

- *Deduplicazione* serve per eliminare l'informazione ridondante in modo da evitare record duplicati nel *database*
  Ad esmpio, un corso può essere promosso da più di un Dipartimento e allora compare in più pagine Web e potrà così essere estratto più volte, anche se il suo contenuto informativo dovrebbe produrre un unico record nel nostro *database*
  Problemi non molto differenti possono insorgere se, ad esempio, in una base riferita ad annate di giornali, si ha RICE, come "Condoleezza Rice" per "Segretario di Stato americano" , ma anche in frasi del tipo "Rice, Wheat, and Beans" riferito a prodotti agricoli. In letteratura si segnala il paradosso che questa operazione ha, in particolare nei *software* commerciali, nomi molto diversi: per chi si occupa di progettazione di basi di dati si chiama "record linkage" oppure "record deduplication"; per chi si occupa di *natural language processing* è nota come soluzione di co-referenze o di anafora resolution; ancora viene chiamata "identity uncertainty" o "object correspondence". Naturalmente ogni campo affronta il problema in maniera leggermente differente, ma fondamentalmente si tratta dello stesso problema

Information Extraction: … (7)



Andrew McCallum, *op. cit.*