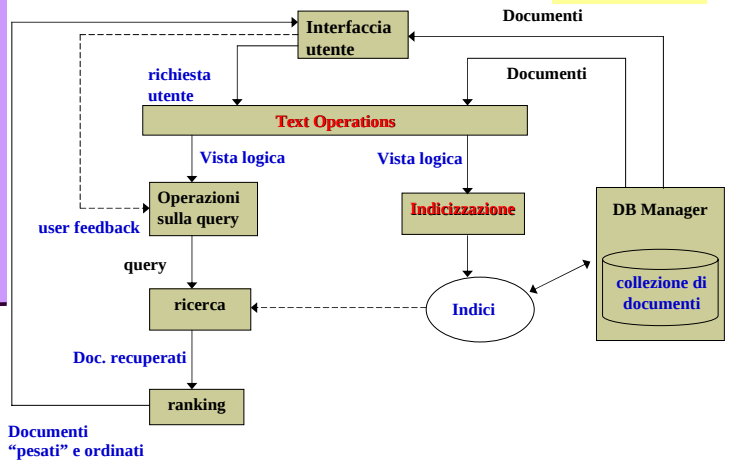
Operazioni sui testi

Architettura di un IRS



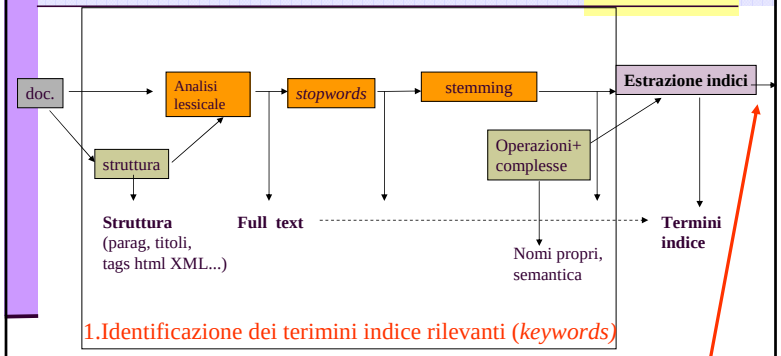
Operazioni sui testi

- Obiettivo:
  - Far in modo che le operazioni di recupero possano essere definite in modo semplice e chiaro, ed implementate in modo efficace ed efficiente
- Trasformano il testo di un documento (o di una query) in una forma standardizzata e compatta
  - Risultato = vista logica del documento (o della query)
  - La vista logica di un documento si basa su un insieme di termini indice estratti dal documento

Operazioni sui testi

- Operazioni principali:
  - Tokenizzazione
  - Analisi lessicale del testo
  - Eliminazione delle "stopword"
  - Lemmatizzazione (stemming)
  - Selezione dei termini caratterizzanti (termini indice)

Operazioni sui testi



Analisi lessicale

- Un testo è trasformato in una lista di parole dette token
- Approccio usuale:
  - Conversione da maiuscolo a minuscolo
  - Ignorare numeri e punteggiatura e considerare solo stringhe contigue di caratteri alfabetici
- Problema: talvolta la punteggiatura, i numeri, e la differenza fra maiuscole e minuscole contengono informazione
  - Mosca vs. mosca
  - Radio 101

Eliminazione delle stopword

- Parole troppo frequenti nei documenti non sono utili per determinare il risultato di una interrogazione
- Dipendono dalla lingua
  - Vector Space Retrieval (VSR) utilizza circa 500 parole inglesi.
  - http://bll.epnet.com/help/ehost/Stop_Words.htm
- Una piccola lista di stopword in Inglese:

a	also	an	and	as	at	be	but
by	can	could	do	for	from	go	
have	he	her	here	his	how		
i	if	in	into	it	its		
my	of	on	or	our	say	she	
that	the	their	there	therefore	they		
this	these	those	through	to	until		
we	what	when	where	which	while	who	with
would	you	your					

Eliminazione delle stopword

- Vantaggi
  - L'eliminazione delle stopword consente di ridurre notevolmente le dimensioni del documento originale
  - Permette di aumentare la precisione delle risposte
- Svantaggi
  - perdita di informazione: si può ridurre il potere di richiamo

to be or not to be

Lemmatizzazione e stemming

- Processo di riduzione di un termine al lemma o radice
  - per riconoscere variazioni morfologiche della stessa parola, o unificare parole che rappresentano concetti simili (derivano dalla stessa radice)

    "comput-er","comput-ational","comput-ation"→ "compute"
- L'analisi morfologica è specifica di ogni lingua, e può essere molto complessa
  - in Italiano è molto più complessa che in Inglese
  - I sistemi di stemming più semplici si limitano ad identificare suffissi e prefissi e ad eliminarli
- Non esiste consenso generale sulle tecniche di normalizzazione
  - Anche in questo caso si può generare perdita di informazioni

Stemming – algoritmo di Porter

- E' una procedura molto semplice
  - iterativamente riconosce ed elimina suffissi e prefissi noti senza utilizzare un dizionario (lemmario)
- Può generare termini che non appartengono alla lingua:
  - "computer", "computational", "computation" → **"comput"**
- Errori di "raggruppamento" (unifica parole con significato differente)
  - organization, organ → organ
  - arm, army → arm
- Errori di "omissione" (non riconosce varianti morfologiche)
  - cylinder, cylindrical

Operazioni più complesse sui testi

- Analisi di nomi propri, date, espressioni monetarie e numeriche
  - La città di **Mosca**, il compositore **Verdi**
  - **April 15th, 2003**, **15-4-03**…
  - **5 millions euros**
- Riconoscimento dei gruppi di nomi
  - Un gruppo di nomi è un insieme di nomi la cui distanza nel testo non supera una data soglia (es: 3)
  - Es: Medical Instrument inc, Consiglio di amministrazione, week end, ..
- Analisi della struttura del testo
  - Es: I termini nel titolo o nei paragrafi hanno un peso maggiore, i termini in grassetto o sottolineati..
- Analisi semantica: associare a parole singole o a porzioni di testo dei concetti di una ontologia
  - "L'albergo dispone di una ampia piscina…" piscina $\xrightarrow{\text{is\_a}}$ hotel_facility
  - "L'albergo si trova in Val Badia…" Val_Badia $\xrightarrow{\text{is\_in}}$ Dolomiti ..

Esempio

- Si supponga di avere il seguente testo:

```
Documents are an interesting application
field for data mining techniques
```

- Effettuiamo solo le seguenti operazioni:
  1. Analisi lessicale
  2. Stemming
  3. Rimozione di stopwords

Esempio: analisi lessicale e stemming

```
(Documents, 1)
(are, 2)
(an, 3)
(interesting, 4)
(application, 5)
(field, 6)
(for, 7)
(data, 8)
(mining, 9)
(techniques, 10)
(., 11)
```

```
(document_N_PL, 1)
(be_V_PRES_PL, 2)
(an_DET, 3)
(interesting_A, 4)
(application_N_SG, 5)
(field_N_SG, 6)
(for_PP, 7)
(data_N_SG, 8)
(mining_N_SG, 9)
(technique_N_PL, 10)
(STOP, 11)
```

Si è aggiunta informazione morfologica:

N = noun, PL = plural, V = verb, PRES = present form,
DET = determinant, A = adjective, SG = singular,
PP = preposition

Esempio: eliminazione stopword e selezione di nomi e aggettivi

```
(document_N_PL, 1)
(be_V_PRES_PL, 2)
(an_DET, 3)
(interesting_A_POS, 4)
(application_N_SG, 5)
(field_N_SG, 6)
(for_PP, 7)
(data_N_SG, 8)
(mining_N_SG, 9)
(technique_N_PL, 10)
(STOP, 11)
```

```
(document_N_PL, 1)

(interesting_A_POS, 4)
(application_N_SG, 5)
(field_N_SG, 6)

(data_N_SG, 8)
(mining_N_SG, 9)
(technique_N_PL, 10)
```