Università di Padova
Dipartimenti di Studi Linguistici e Letterari

Tecnologie per la Traduzione 2020/2021

Reperimento dell'Informazione

Giorgio Maria Di Nunzio

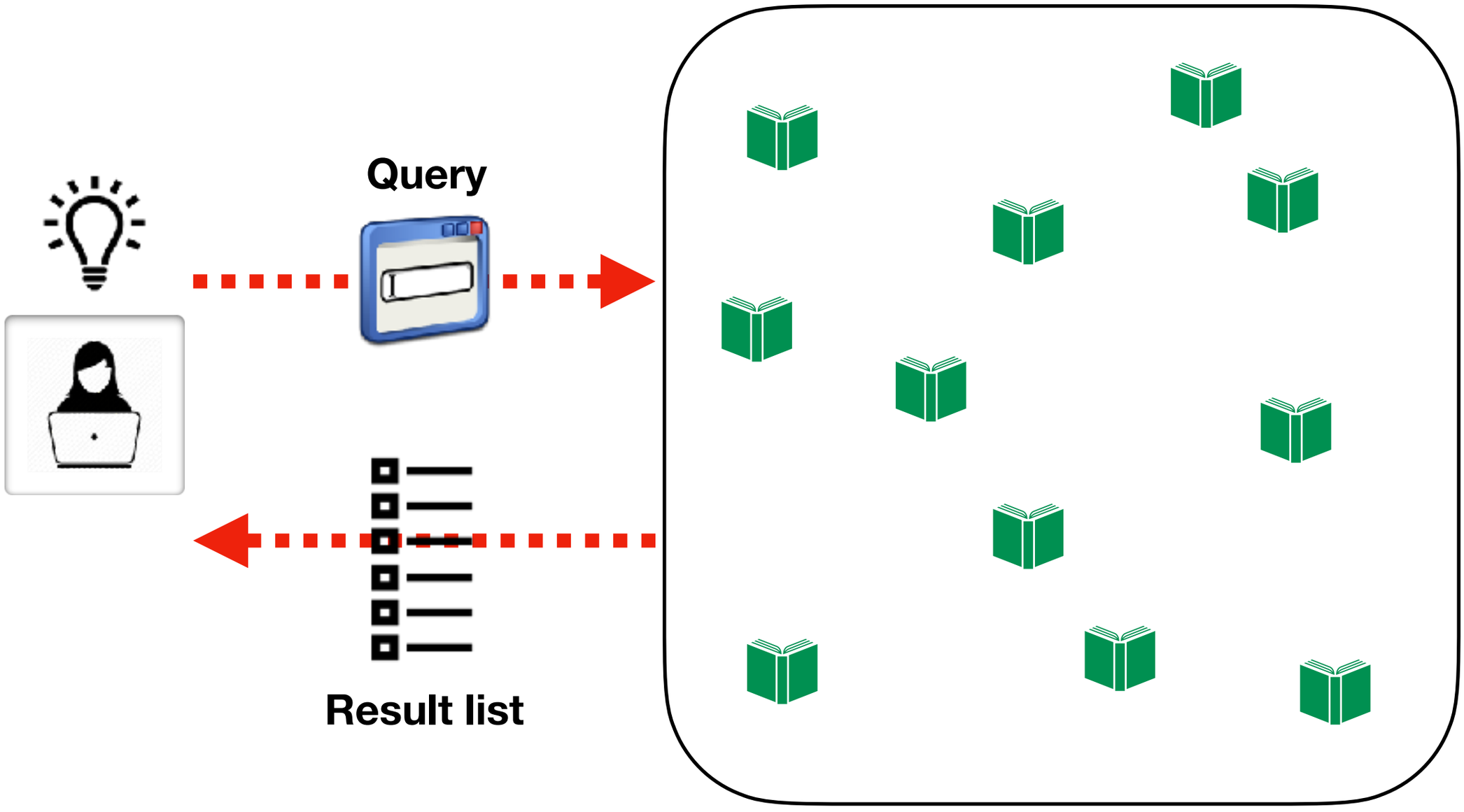
Obiettivi

- Definizione del problema
- Motori di ricerca (anche su Web)
- Valutazione dei risultati di un motore di ricerca

Reperimento dell’informazione

- Il termine reperimento dell’informazione identifica tutte quelle attività utilizzate per scegliere, da una data collezione di documenti, quei documenti che risultano di interesse in relazione ad una specifica esigenza informativa.

Reperimento dell’informazione



Esigenza informativa, interrogazione, risultati

- L’utente ha una specifica esigenza informativa.

- L’utente esprime questa esigenza attraverso una sintesi, una *query* in gergo tecnico, inoltrando la richiesta ad un sistema di ricerca.

- Il sistema interpreta la richiesta e “tenta”, vista l’informazione parziale, di recuperare i documenti più rilevanti.

Reperimento dell’informazione

- Finalità:
  - selezione dei documenti che sono verosimilmente rilevanti per le esigenze informative dell’utente
- Tipologie:
  - interrogazione in forma testuale (non solo)
  - navigazione ipertesti (spesso combinate)
- Modalità:
  - sotto forma di ciclo presentazione/valutazione

Utente vs Sistema

- L’utente gioca un ruolo determinante, l’efficacia del recupero dipende da:
  - quanto l’utente sa cosa sta cercando,
  - come l’utente esprime le sue esigenze informative,
  - la capacità dell’utente di valutare la pertinenza dei documenti ritrovati e di riformulare le interrogazioni.

Dove sono le informazioni?

- Problema della disponibilità di informazioni.
- Gli utenti dovrebbero poter sapere
  - Quali informazioni sono disponibili, ovvero se le informazioni a loro necessarie sono presenti.
  - Come raggiungere le informazioni disponibili.
- La "catalogazione" descrive, in maniera sintetica e di rapido accesso, il contenuto informativo dei documenti presenti in una collezione.

Catalogazione manuale o automatica?

- Il problema di come reperire le informazioni aumenta con la mole dei dati messi a disposizione

- Ad esempio, nelle biblioteche esiste un indice catalografico ed un ordinamento per argomento, autore, ecc.

- La catalogazione viene di solito svolta manualmente ed è quindi molto lunga (e passibile di errori).

- L'informatica consente di automatizzare l'organizzazione dei dati e il loro reperimento.

Indicizzazione automatica

- E’ possibile automatizzare l’estrazione del contenuto informativo, operazione che viene definita indicizzazione.
    - Creare un modello che consenta di estrarre le informazioni rilevanti in modo automatico.
- Nei documenti testuali l’informazione è contenuta nella parole che compongono i documenti.
    - E’ più difficile definire il contenuto semantico di documenti in formati non testuali, ad esempio musica o immagini.
- Una volta indicizzati i documenti, è possibile effettuare delle ricerche negli indici dei documenti.
- La ricerca negli indici richiede meno operazioni (è quindi più efficiente)

Reperimento dell’informazione sul Web

- Il Web è una collezione di documenti ma non è gestito in maniera unitaria e/o controllata
    - Chiunque può creare una pagina o un sito Web, nel quale può mettere qualsiasi informazione.
    - Non è possibile effettuare controlli sul contenuto dei siti e sulla loro attendibilità.
- Ogni giorno vengono creati centinaia di nuovi siti, e altrettanti siti scompaiono.
- Impossibile stabilire il numero di pagine Web (centinaia di miliardi?).
- E’ impossibile per un utente tener traccia di questa continua evoluzione senza l’ausilio di strumenti informatici
    - Il problema principale è scoprire dove si trova l’informazione.

Motori di ricerca (su Web)

- La rapida espansione del Web ha reso necessaria la scrittura di programmi che aiutino l'utente a reperire l'informazione.

- Un motore di ricerca è un sistema di programmi per il reperimento dell'informazione.

- Un motore di ricerca del Web è un sistema in grado di
  - individuare,
  - indicizzare e
  - reperire le pagine Web

Individuazione delle pagine Web

- Il Web Search Agent (WSA anche chiamato *crawler*, o *spider*) è quel componente di un motore di ricerca che viene impiegato per l'individuazione delle pagine Web.

- Il WSA localizza le pagine Web, e in generale tutti documenti accessibili sui server Web, lavorando *ricorsivamente*

  - Si parte da una lista di URL noti

  - Si analizzano i documenti per vedere se ci sono link a nuovi URL al di fuori della lista

  - Si aggiorna la lista di URL e visita i documenti agli URL aggiunti

  - Si ripete fino a che non si sono esplorati tutti i link

Aggiornamento della lista delle pagine Web

- Il processo automatico di individuazione delle pagine Web permette di fare una “copia” dei documenti di un sito Web in un certo istante di tempo.
  - Vedi Web Archive
- Ciò dignifica che le pagine che ha individuato il WSA diventeranno presto “vecchie” rispetto alla versione attuale.
- Per questo motivo, i motori di ricerca “navigano” costantemente il Web per raggiornare la collezione di pagine
  - Vedi Google Regular Crawling

Dall’individuazione all’indicizzazione

- “E’ possibile automatizzare l’estrazione del contenuto informativo, operazione che viene definita indicizzazione.”

- Una volta raccolta (individuata) la collezione di documenti, è necessario indicizzarla per permettere una ricerca più efficiente.

    - Quello che si vuole evitare è leggere “tutti” i documenti per poter ottenere l’informazione che si sta cercando.

- Un indice è un’organizzazione dei dati che permette all’utente di trovare rapidamente l’informazione che sta cercando.

Indicizzazione automatica

Analisi lessicale

Rimozione parole comuni

Lemmatizzazione o Stemming

Pesatura

Creazione indice

Esempio

- 3 documenti
  - Doc1: “shipment of gold damaged in a fire”
  - Doc2: “delivery of silver arrived in a silver truck”
  - Doc3: “shipment of gold is arriving in trucks”

Analisi lessicale

- Separare le parole (segmentazione?)
  - Doc1: “shipment” “of” “gold” “damaged” “in” “a” “fire”
  - Doc2: “delivery” “of” “silver” “arrived” “in” “a” “silver” “truck”
  - Doc3: “shipment” “of” “gold” “is” “arriving” “in” “trucks”

Rimozione parole comuni (stop word list)

- Eliminare parole comuni (“of”, “in”, “a”, ...)
- Doc1: “shipment” “gold” “damaged” “fire”
- Doc2: “delivery” “silver” “arrived” “silver” “truck”
- Doc3: “shipment” “gold” “arriving” “trucks”

Lemmatizzazione o stemming

- Esempio di stemming:
  - D1: “shipment” “gold” “damag” “fir”
  - D2: “deliveri” “silver” “arriv” “silver” “truck”
  - D3: “shipment” “gold” “arriv” “truck”

Pesatura

- Per ogni documento, conto le occorrenze di ogni termine:
  - Doc1: (“shipment”, 1) (“gold”, 1) (“damag”, 1) (“fir”, 1)
  - Doc2: (“deliveri”, 1) (“silver”, 2) (“arriv”, 1) (“truck”, 1)
  - Doc3: (“shipment”, 1) (“gold”, 1) (“arriv”, 1) (“truck”, 1)

Creazione indice (index)

	arriv	damag	deliveri	fir	gold	silver	shipment	truck
Doc1	0	1	0	1	1	0	1	0
Doc2	1	0	1	0	0	2	0	1
Doc3	1	0	0	0	1	0	1	1

Creazione indice (inverted index)

	Doc1	Doc2	Doc3
arriv	0	1	1
damag	1	0	0
deliveri	0	1	0
fir	1	0	0
gold	1	0	1
silver	0	2	0
shipment	1	0	1
truck	0	1	1

Modelli di recupero

- L’obiettivo è recuperare i documenti che sono “probabilmente” rilevanti rispetto all’interrogazione dell’utente.

- Vi sono vari modelli di recupero, che possono essere suddivisi in due grandi famiglie:

  - exact match: vengono individuati in modo esatto i documenti che soddisfano l'interrogazione e quelli che non la soddisfano.

  - best match: viene effettuata una stima della rilevanza di un documento ad una data interrogazione. I documenti vengono ordinati per una misura di similarità con l'interrogazione e sono proposti quelli sopra una prefissata soglia.

Exact match: modello Booleano

- I termini dell'interrogazione sono collegati mediante operatori logici (AND, OR, NOT). Ad esempio, cerco di documenti che:

  - contengono il termine t1 e il termine t2: **t1 AND t2**
  - contengono t1 o non contengono il termine t2: **t1 OR (NOT t2)**

- Vengono recuperati solo i documenti che soddisfano esattamente l'espressione richiesta.

- Alcuni sistemi permettono l'utilizzo operatori di prossimità (termine t1 "vicino" a t2) o troncamento dei termini.

Exact match: modello Booleano

- Richiede (un minimo) di conoscenza dell’algebra Booleana.
  - Poco “user friendly”.
- Mancanza di controllo sulla dimensione dell’insieme dei documenti recuperati.
- Non è possibile l’ordinamento dei documenti.
  - Sono tutti importanti allo stesso modo.
- Non è possibile pesare i termini (un termine è presente o no).

Best match: Tf-Idf

- Non tutte le parole di un documento ne descrivono il contenuto con la stessa “quantità di informazione”.

- Questa quantità si può misurare associando un “peso” che ne quantifica l’importanza.

- Il peso di un termine tiene normalmente conto della frequenza del termine nel documento e nella collezione.

Best match: Tf-Idf

- Term Frequency (Tf)
  - Frequeza di una parola all’interno di un documento (o di una query)
  - Quante volte appare la parola?
  - Quante volte appare la parola rispetto al numero totale di parole presenti nel documento?
- Document Frequency (Df)
  - Frequenza di una parola all’interno di una collezione di documenti
  - In quanti documenti appare la parola?
  - In quanti documenti appare la parola, rispetto al numero totale di documenti della collezione?
- Inverse Document Frequency (Idf) = 1/df

Rilevanza di un documento

- La bontà di un sistema di reperimento dipende da quanti documenti reperiti sono effettivamente rilevanti per l'esigenza informativa dell'utente.

- Le prestazioni di un motore di ricerca potrebbero essere calcolate se si conoscesse l'insieme totale dei documenti che rispondono alle esigenze informative dell'utente.

  - E' praticamente impossibile conoscere la rilevanza di milioni di documenti, o miliardi se ci si riferisce al Web

  - La rilevanza è soggettiva e può variare nel tempo

  - Il giudizio sulla rilevanza di un documento influisce sul giudizio dei successivi

Valutazione

- E’ auspicabile che un sistema per il reperimento delle informazioni presenti tutti e soli i documenti rilevanti per l’utente.

- Se così fosse, l’utente non avrebbe bisogno di valutare i documenti, e la ricerca si esaurirebbe in un’unica iterazione.

- Vi sono due possibili comportamenti negativi, che rendono difficile la valutazione e oneroso il reperimento

- Effetto rumore

    - Il sistema reperisce anche documenti non rilevanti; la valutazione e la consultazione sono più onerose perché i documenti rilevanti sono diluiti

- Effetto silenzio

    - Il sistema non reperisce alcuni documenti che sarebbero invece rilevanti; l’utente non può accedere ad una parte dell’informazione

Relevance Feedback - Query Refinement

- Vedi Capitolo 9 “Relevance feedback & Query expansion” del libro “Introduction to Information Retrieval”
  Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.