

# Information Retrieval

- **Information Retrieval**
- **Probabilistic Information Retrieval**
- **Criteri di valutazione di un sistema di IR**

# Fabbisogno informativo

- 1. Risposta puntuale ad una richiesta**
  - *ES: “Chi eroga corsi di formazione sulle BSC?”*
- 2. Produzione di raccolte**
  - *ES: “Case histories delle BSC in istituzioni finanziarie”*
- 3. Supporto alla elaborazione di studi**
  - *ES: “Metodologie e sistemi di metriche per l’ analisi della performance”*

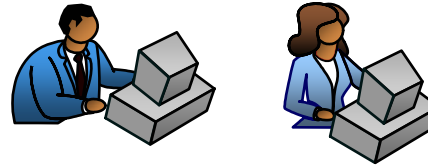
# Information Retrieval (IR)

- La definizione “information retrieval” è stata coniata da Calvin Mooers nel 1952
- Obiettivo dell’IR è di recuperare, all’interno di un insieme di documenti , **tutti e solo** i documenti **rilevanti** per un particolare utente con una particolare richiesta informativa
- Legge di Mooers : “Un sistema di reperimento delle informazioni tenderà a non essere usato quando trovare le informazioni è “more painful and troublesome” ( “più noioso e doloroso”) che non trovarle”

# Information retrieval

- **Data retrieval:** trovare oggetti che soddisfino condizioni chiaramente specificate mediante una espressione regolare o di algebra relazionale.
  
- **Information retrieval:**
  - Trovare informazioni significative per l'utente che effettua la ricerca.
  - Richiede una *interpretazione* della richiesta dell'utente
  - I documenti recuperati non possono essere classificati *tout court* come "giusti" o "sbagliati", ma vanno associati ad una misura di *rilevanza* rispetto alla richiesta dell'utente
  - La nozione di rilevanza è basilare nei sistemi di Information Retrieval

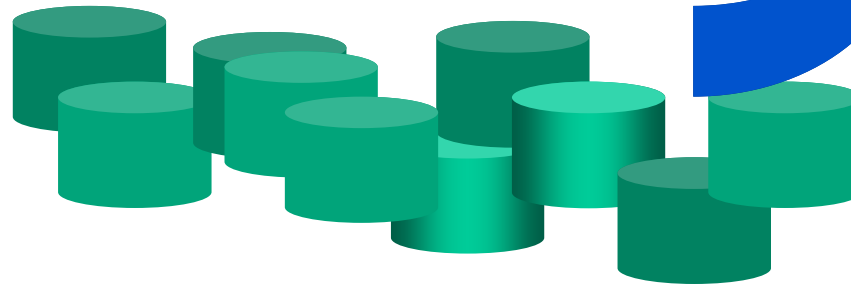
# Information Retrieval



Processo di  
formulazione della richiesta  
(query)

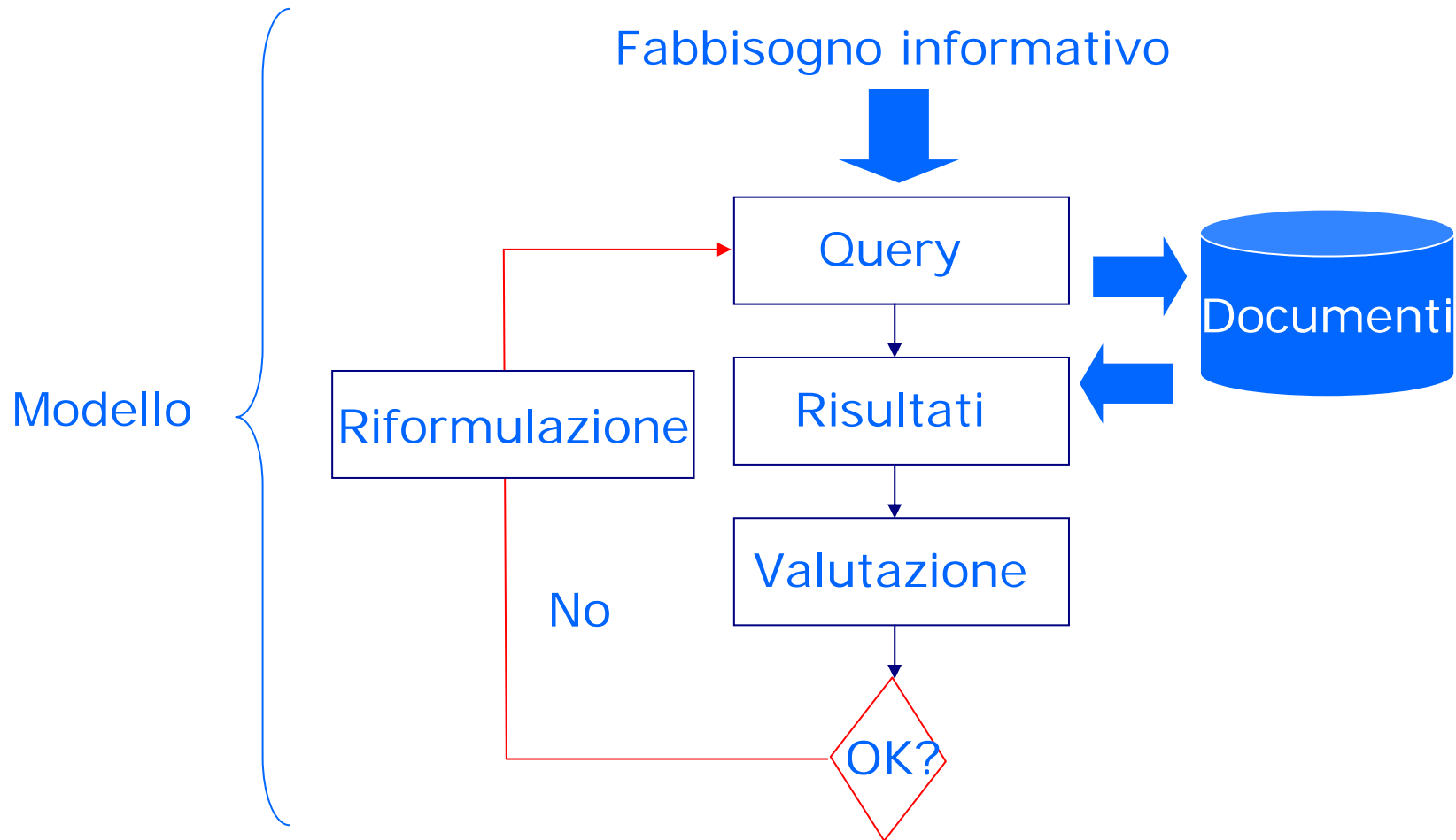
Processo di  
Identificazione del contenuto  
(indicizzazione)

Corrispondenza tra  
rappresentazione della  
domanda  
e rappresentazione del  
documento



- **Da un punto di vista formale, un sistema di Information Retrieval è rappresentabile (Maron – Cooper) da una “quadrupla” composta da :**
  - un insieme di rappresentazioni logiche dei documenti  $D$ ,
  - un insieme di rappresentazioni logiche dei fabbisogni informativi  $Q$
  - il modello di riferimento per la rappresentazione dei documenti e delle richieste e delle relazioni tra loro,  $F$
  - una funzione di “ranking” cioè una funzione in base alla quale ordinare i documenti recuperati per soddisfare la richiesta  $F(Q_i, D_j)$

# Processo di ricerca e recupero dell'informazione: modello statico

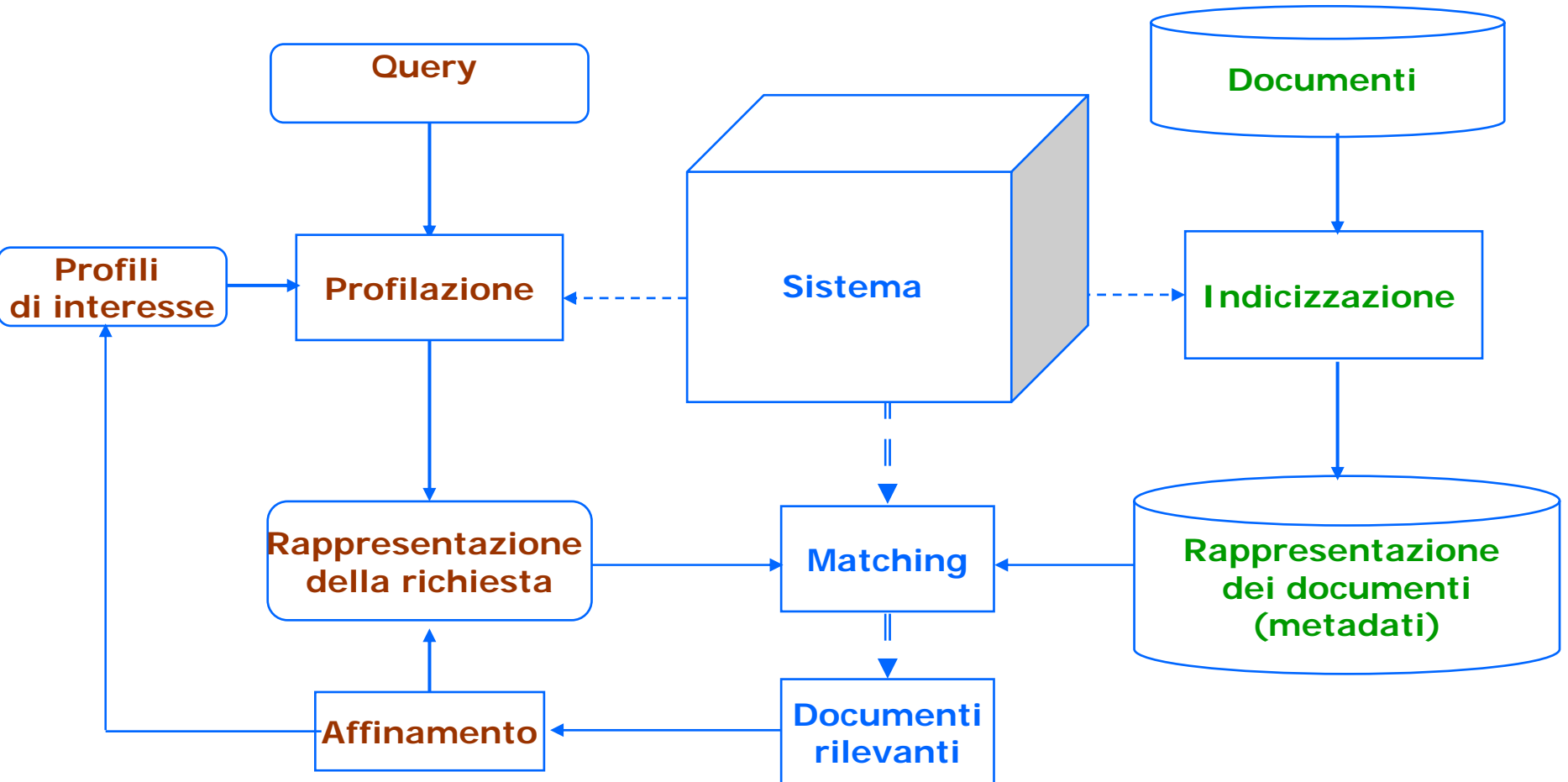


## Limiti del modello statico

- Nuove informazioni possono portare a nuove idee e a nuove direzioni di ricerca (il fabbisogno informativo si modifica)
- Il bisogno informativo non segue una logica binaria (si/no) ma viene soddisfatto attraverso una serie di scelte e elementi di informazione recuperati progressivamente nel corso della ricerca

# Probabilistic Information Retrieval

# Struttura del sistema IR



# Ricerche concettuali basate su analisi statistica e ranking probabilistico

- L' identificazione dei concetti si basa sulla teoria dell'informazione di Shannon
- Ranking dei documenti basato su modelli probabilistici (statistica soggettiva basata su modelli bayesiani).
- Vantaggi : Alta precisione e recall . Utilissima per ricerche su contesti non conosciuti

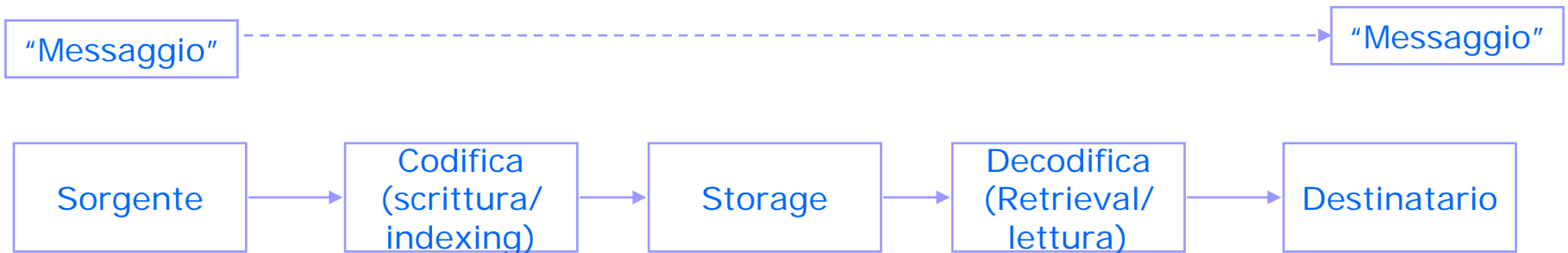
# Le tappe fondamentali

1950 - 1960	Analisi statistica del linguaggio	Shannon, Luhn,...
1960 - 1970	Modelli probabilistici	Maron, Kuhns
1970 - 1980	Espansione della query	Salton, Rocco,...
1980 - 1990	Relevance Feedback	Robertson, Sparck-Jones
1990 -	Latent Semantic Analysis	Dumais, Deerwester

# Identificazione dei concetti : teoria dell'informazione di Shannon

Secondo Shannon, il contenuto informativo di un "messaggio" è rappresentato dalla sua probabilità di presentarsi in un insieme di messaggi possibili: maggiore è la probabilità di realizzarsi, minore è il contenuto informativo.

È abbastanza intuitivo che sarà il messaggio meno probabile a portare la massima quantità di informazione quando si presenta.



Shannon- Weaver Teoria matematica delle comunicazioni 1983

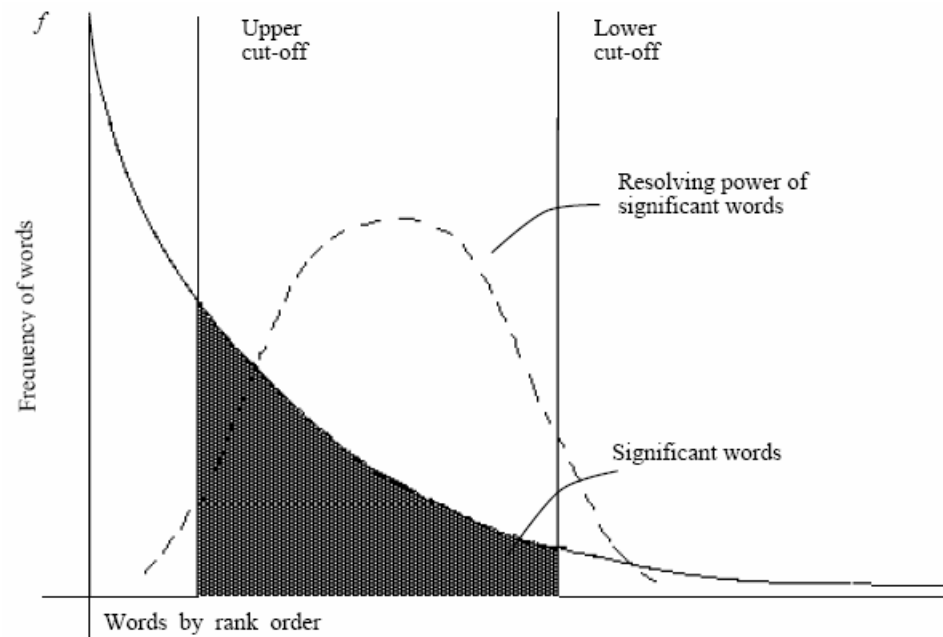
# Formulazione

- Dato un insieme finito A (parole o frasi) costituito da n unità (lettere di un predefinito alfabeto) e definito P il suo vettore di probabilità:

$$H(P) \doteq - \sum_i p_i \log_2 p_i \equiv \sum_i p_i \log_2 \frac{1}{p_i}$$

- La funzione H identifica ciò che Shannon definì l' entropia del messaggio, cioè la misura della sua indeterminatezza.

- Legge di Luhn: in estrema sintesi la legge di Luhn stabilisce che la capacità dei termini di discriminare il contenuto dei documenti è massima nella posizione intermedia tra i 2 livelli di cut-off della distribuzione di frequenza.



# Inferenza bayesiana

- In termini semplici il teorema di Bayes parte dalla considerazione che non tutti gli eventi siano direttamente osservabili.
- Se non possiamo osservare l'evento A la probabilità " a priori" di A sarà  $P(A)$
- Però se un evento A è in qualche modo connesso ad un evento B che possiamo osservare, la probabilità a posteriori di A sarà  $P(A|B)$ .
- Il teorema di Bayes stabilisce come calcolare la probabilità " a posteriori" di A a partire dalla probabilità, nota, di B.

$$P(A_i | B) = \frac{P(B | A_i) * P(A_i)}{\sum_k P(B | A_k) * P(A_k)}$$

- In termini di Information Retrieval, identificato un set di documenti che si sa essere rispondenti alla query, (evento B) è possibile calcolare per gli altri documenti la probabilità che hanno di essere rilevanti.

# Relevance Weighting

Formula di Robertson – Sparck Jones

$$W_i = \log \frac{\left( \frac{r + 0.5}{R - r + 0.5} \right)}{\left( \frac{n - r + 0.5}{N - n - R + r + 0.5} \right)}$$

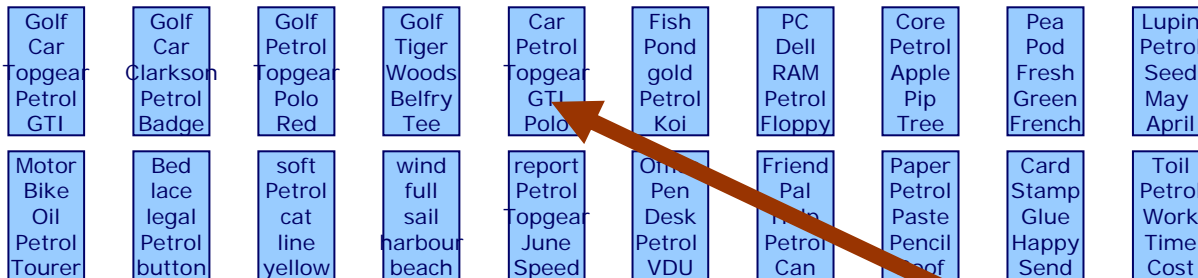
r	Numero di documenti rilevanti che contengono il termine
R	Numero di documenti rilevanti
n	Numero di documenti che contengono il termine
N	Numero di documenti

# Latent Semantic Analysis

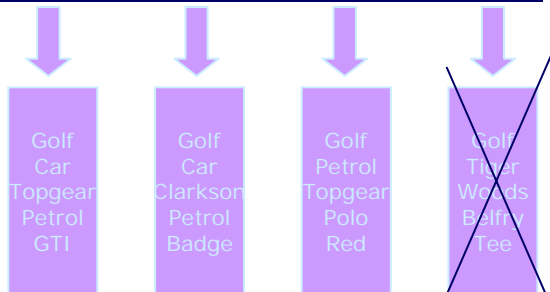
- La capacità di individuare documenti che sono rilevanti, ma non contengono le parole della query
- Analizza la co-occorrenza di termini in un sottoinsieme di documenti (es. i documenti rilevanti per una query)
- Utilizza la tecnica Singular Value Decomposition (SVD) di algebra lineare per l'identificazione dei "related topic"

# Esempio

**Documenti totali: 20**



**Ricerca : Golf : 4 documenti selezionati**



**Relevance ranking**

40      30      30



<b>Golf</b>	1	1	1	1
<b>Car</b>	1	1		
<b>topgear</b>	1		1	
<b>petrol</b>	1	1	1	
<b>GTI</b>	1			
<b>Clakson</b>		1		

**Nuovo documento  
selezionato**

**Related Topic:  
Car, Topgear**

Car  
 $2 * (20/3) = 13$   
 Topgear  
 $2 * (20/3) = 13$   
 Petrol  
 $3 * (20/16) = 4$

**Ranking dei termini  
sui documenti  
selezionati**

# Valutazione di un sistema di Information Retrieval

# Valutazione di un sistema di IR

- Per l'utente il documento rilevante per una query può essere il documento che:
  - risponde precisamente alle esigenze dell'utente
  - risponde parzialmente alle esigenze dell'utente
  - suggerisce una fonte di informazioni
  - fornisce informazioni contestuali sull'argomento di interesse
  - richiama alla memoria dell'utente altre conoscenze

- **Fattori che influenzano la valutazione di rilevanza (Saracevic):**
  - Utente: il contesto dell'interrogazione e i limiti posti all'interrogazione
  - Interrogazione: struttura e classificazione delle interrogazioni
  - Chi effettua la ricerca: esperienza e percorsi cognitivi di chi effettua la ricerca
  - Ricerca: modalità di formulazione della ricerca
  - Documenti recuperati

# Criteri di valutazione

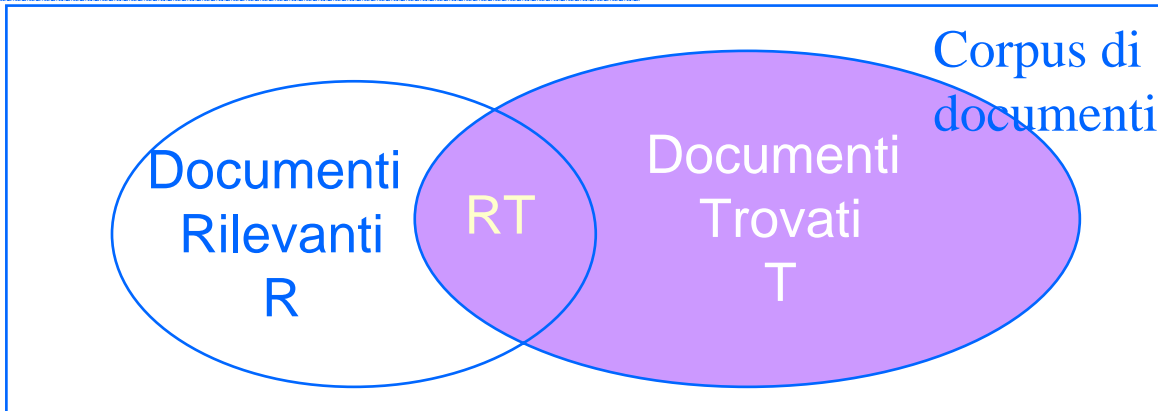
- Una vasta bibliografia è disponibile sulla valutazione dei sistemi di IR.
- Cleverdon definisce 6 criteri di valutazione:

1	The <u>coverage</u> of the collection, that is, the extent to which the system includes relevant matter
2	the <u>time lag</u> , that is, the average interval between the time the search request is made and the time an answer is given
3	the form of <u>presentation</u> of the output
4	the <u>effort</u> involved on the part of the user in obtaining answers to his search requests
5	the <u>recall</u> of the system, that is, the proportion of relevant material actually retrieved in answer to a search request
6	the <u>precision</u> of the system, that is, the proportion of retrieved material that is actually relevant

## Misure di efficacia

RECALL	$rR / (rR + nrR)$	Percentuale di documenti rilevanti recuperati sul totale dei documenti rilevanti presenti nell' insieme
PRECISIONE	$rR / (rR + rNR)$	Percentuale di documenti rilevanti sul totale dei documenti recuperati
SILENZIO	$nrR / (rR + nrR)$	Percentuale di documenti rilevanti non recuperati sul totale dei documenti rilevanti presenti nell' insieme
RUMORE	$rNR / (rR + rNR)$	Percentuale di documenti non rilevanti sul totale dei documenti recuperati

# Efficacia di un sistema di IR



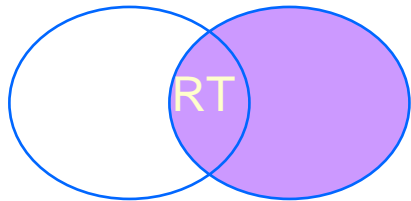
Efficacia

$$\text{Precisione} = RT / T$$

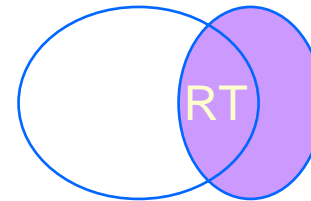
Proporzione di materiale trovato che è rilevante

$$\text{Recall} = RT / R$$

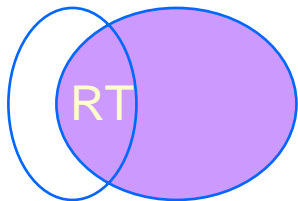
Proporzione di materiale rilevante che viene trovato



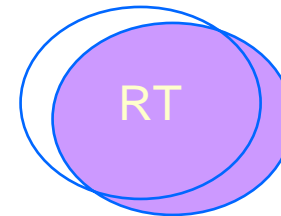
Bassa Precisione  
Basso Recall



Alta Precisione  
Basso Recall



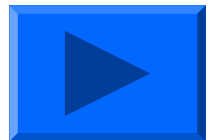
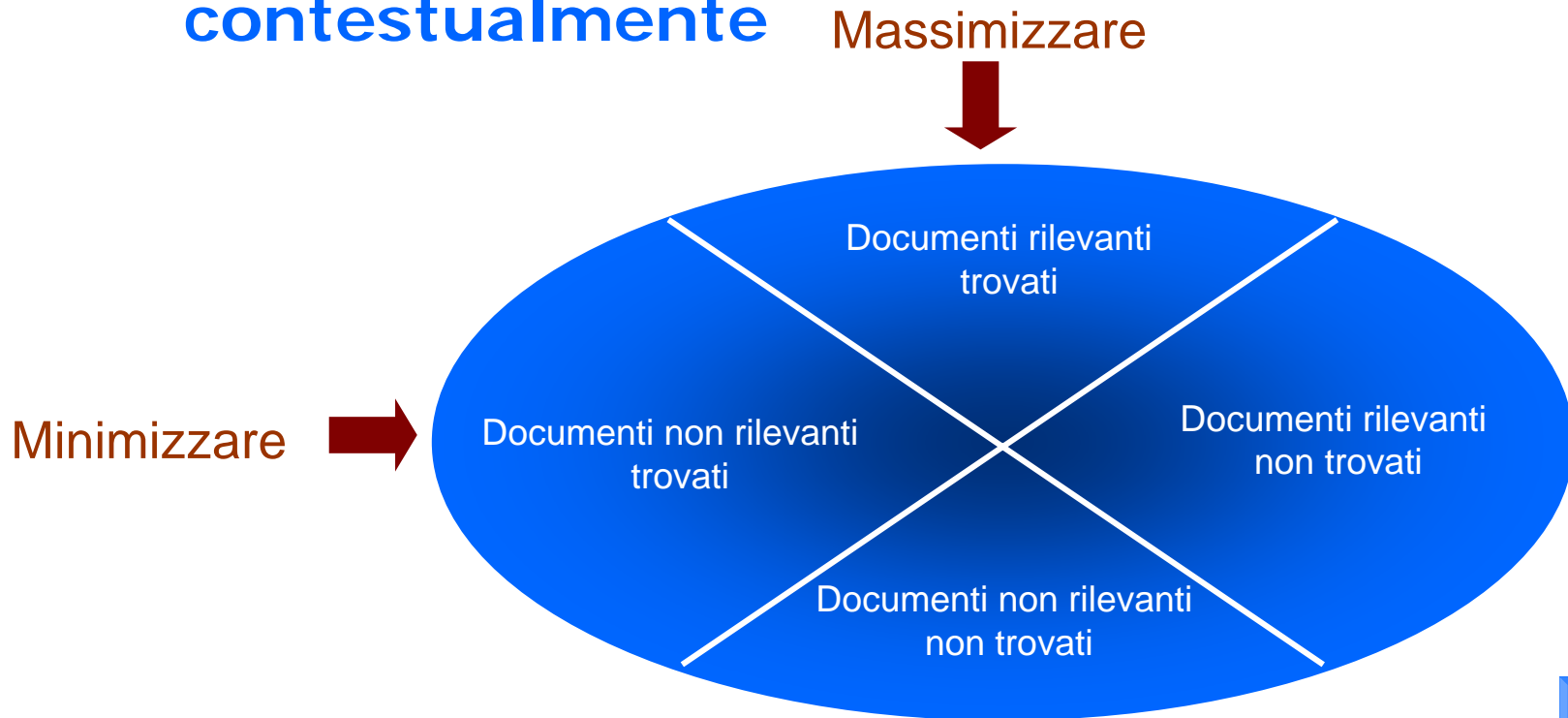
Bassa Precisione  
Alto Recall



Alta Precisione  
Alto Recall

# Precisione e richiamo

- I 2 parametri vanno considerati contestualmente



Parametro	Modelli statistico-probabilistici
Precisione e recall	Come dimostrano i molti test disponibili, grazie alla integrazione di Pattern Matching, Relevance Feedback e LSA, queste tecnologie riescono a fornire contemporaneamente alto recall e alta precisione
Contesti di Applicabilità	Essendo indipendenti da una specifica lingua o gergo specialistico sono applicabili in ogni contesto
Facilità di ricerca da parte dell' utente	L'utente può formulare la sua query in linguaggio naturale e non ha bisogno di alcuna conoscenza della sintassi del motore di ricerca
Facilità di analisi dei risultati	I risultati possono essere presentati secondo ordinamenti diversi tra cui quello della <u>rilevanza</u> : i documenti più rilevanti sono i primi nella lista sottoposta all' utente
Semplicità di implementazione	Grazie all' elevato grado di automazione dei processi di indicizzazione e non richiedendo la predisposizione di appositi dizionari questi sistemi sono rapidi e semplici da implementare