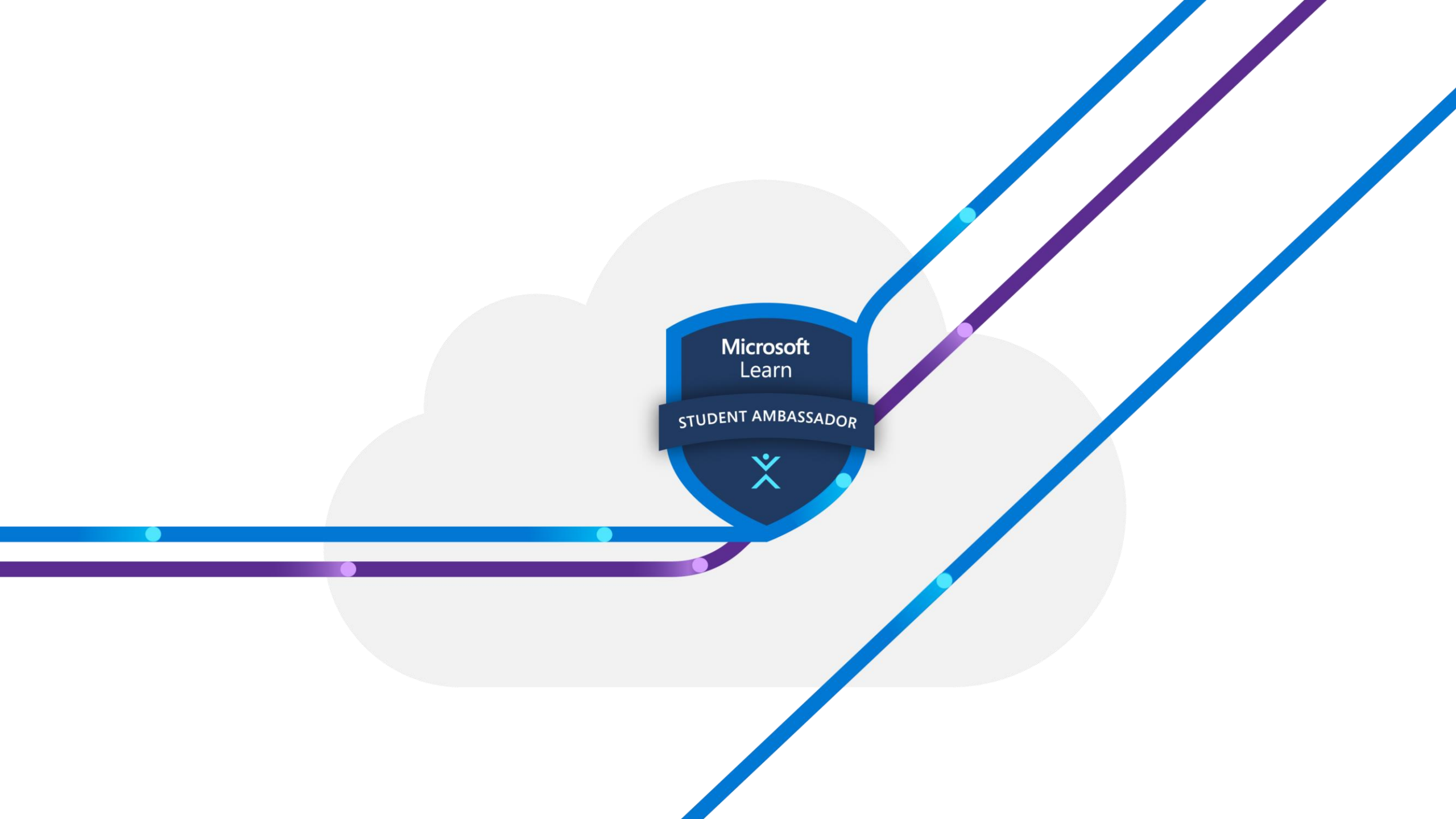

Microsoft
Learn
STUDENT AMBASSADOR

Microsoft Learn
Student Ambassadors

Web search – PageRank & HITS

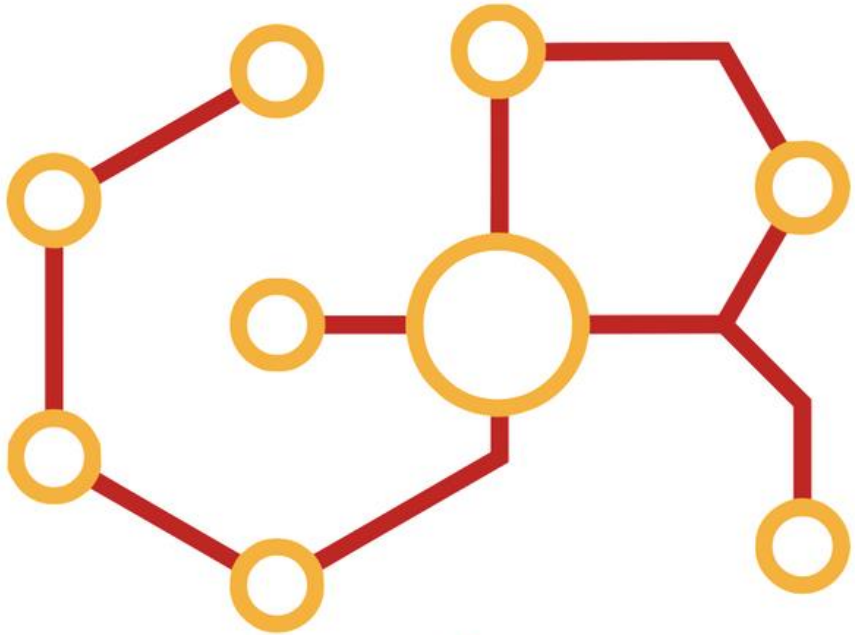
Mario Cuomo

20.01.2022

Microsoft Learn
Student Ambassadors

MARIO CUOMO

mariocuomo.github.io
linkedin/in/mariocuomo
@mariocuomo.exe
@mariocuomoEXE

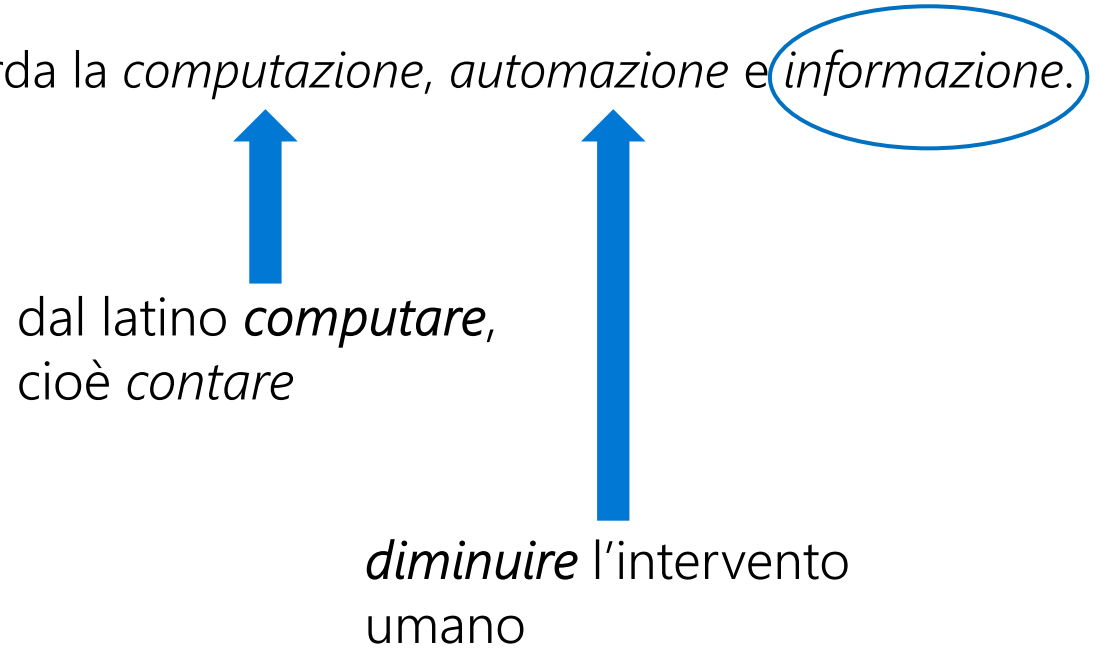


GraphRM

Meetup #AperiTech

Information Retrieval

Branca della ***computer science*** – tutto ciò che riguarda la *computazione*, *automazione* e *informazione*.



Information Retrieval

Gerard Salton,
the father of IR

Information Retrieval (*IR*)

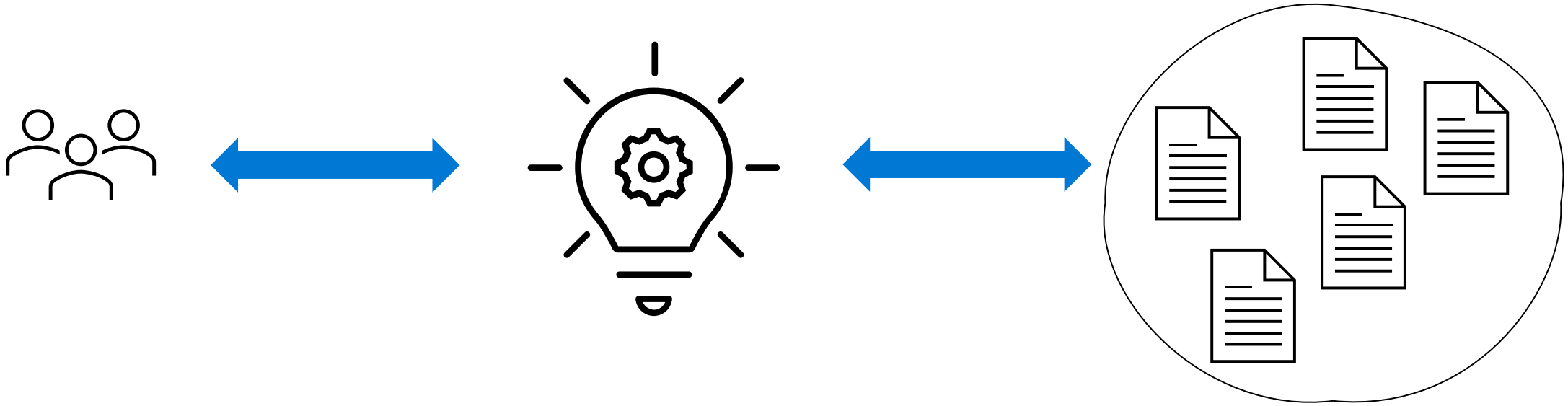
Salton, *Automatic Information Organization and Retrieval*, 1968

Information Retrieval: tutto ciò che riguarda la *struttura*, *analisi*, *organizzazione*, *memorizzazione* e *ricerca* dell'*Informazione*

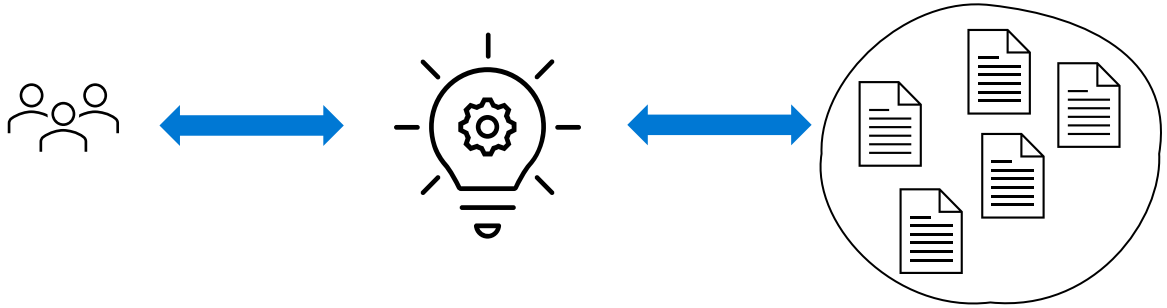
IERI $informazione = documento\ testuale$

OGGI $informazione = qualsiasi\ risorsa\ informativa$

Task *IR*



Task *IR*



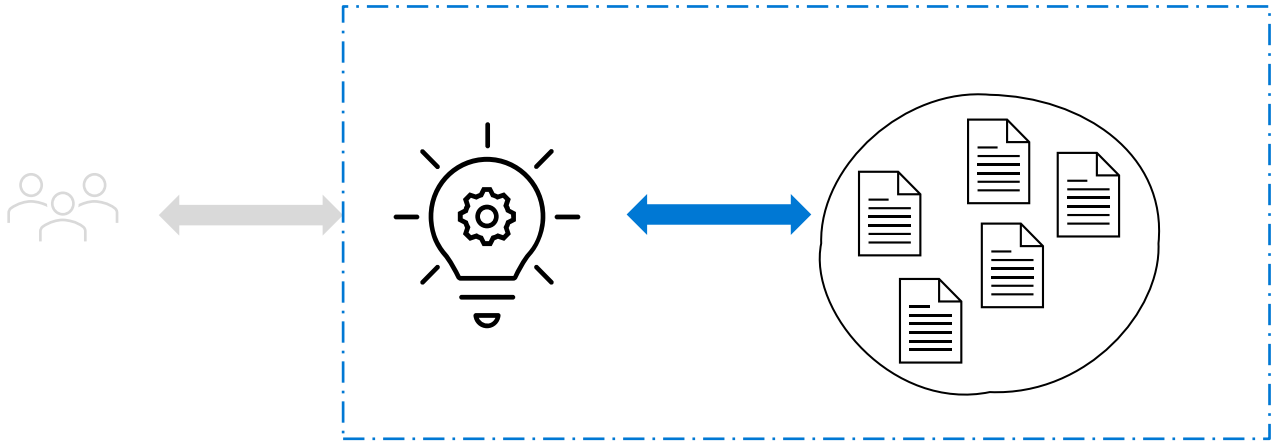
- $\boldsymbol{ricerca\ ad-hoc}$: recuperare documenti *rilevanti* a fronte di una *query*
- $\boldsymbol{filtraggio}$: recuperare documenti *rilevanti* a fronte di una *query* e un *modello utente*
- $\boldsymbol{classificazione}$: assegnare *etichette rilevanti* ai documenti in base a una *caratteristica*
- $\boldsymbol{question-answering}$: fornire *risposta* a una *domanda* in *formato naturale*

Applicazione *IR – Motori di Ricerca*

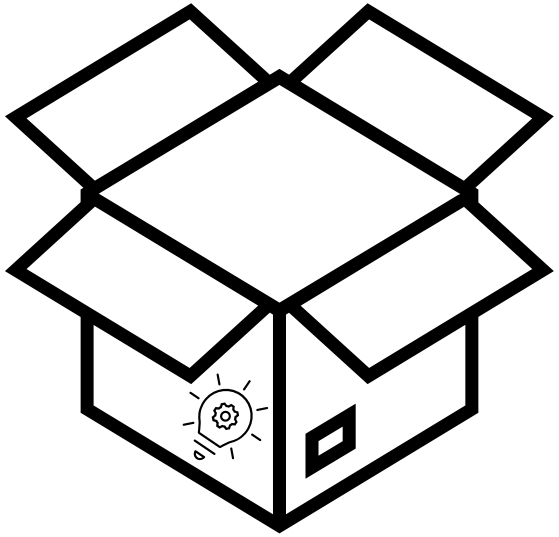
Diversi tipi di motori di ricerca

- ***Web search***
- *Vertical search*
- *Enterprise search*
- *Desktop search*
- *Peer-to-peer search*

Architettura di un *Web Search Engine*



Processo di *indicizzazione*



- ***Text Acquisition***
- ***Text Transformation***
- ***Data Store***
- ***Index Creation***

Modulo *Text Acquisition*

Si ha un agente software – un ***crawler*** – che naviga per il web con l'obiettivo di *visitare* e *memorizzare* le pagine web.

Modulo *Text Transformation*

I documenti sono elaborati prima di essere memorizzati

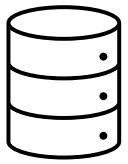
- *Parsing/Tokenizzazione*
- *Stopping*
- *Stemming*

Si identificano gli *index term*

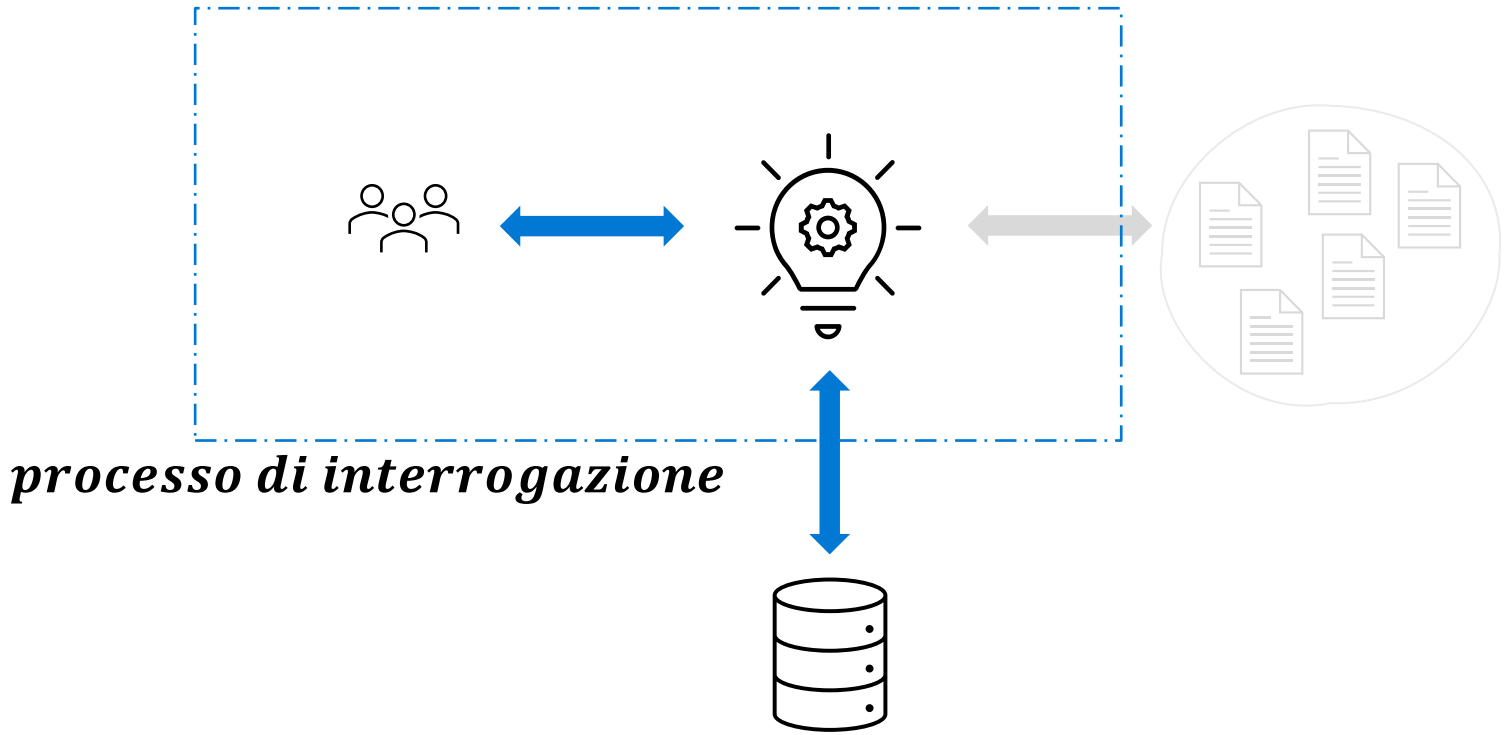
Modulo *Index Creation*

Core del processo di indicizzazione.

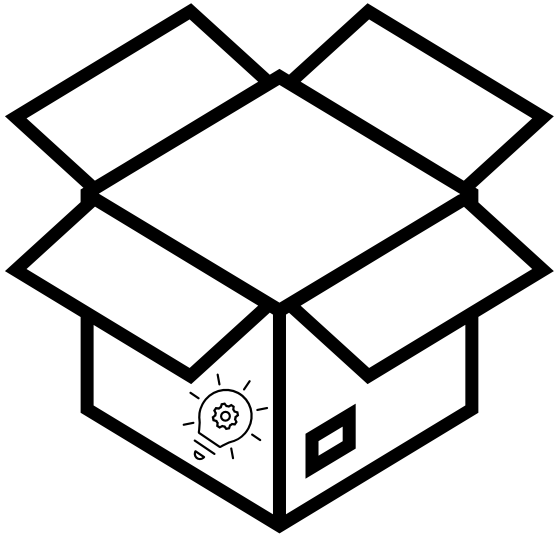
Trasforma l'informazione da *Document→Terms a Term→Documents*

Crea un ***Indice*** 

Architettura di un *Web Search Engine*



Processo di *interrogazione*



- ***User Interaction***
- ***Ranking***
- ***Evaluation***

Modulo *User Interaction*

L'utente sottomette la query che può essere pre-elaborata.

- *Tokenizzazione/Stopping/Stemming*
- *Spell checking e query Suggestion*
- *Query exapansion*

Modulo *Ranking*

Si restituiscono i documenti *rilevanti* per la query.

I documenti sono *ordinati*.

PAGERANK

Algoritmo utilizzato per trovare un metadato – il ***rank*** – delle pagine web.

1996 – Larry Page e Sergey Brin.
Utilizzato da *Google Search Engine*.

ASSUNZIONE
Una pagina web è *autorevole* se ha molti link in ingresso

Attenzione alle *link farm* malevole

PAGERANK

Algoritmo *agnostico* rispetto al contenuto.

Si ha un ***random surfer*** che naviga tra le pagine web

2 operazioni possibili – con diversa probabilità di scelta

- spostarsi su un link a caso tra quelli all'interno della pagina
- spostarsi su una pagina a caso tra quelle già visitate

PAGERANK

$$PR(u) = \frac{\lambda}{N} + (1 - \lambda) \sum_{v \in B_u} \frac{PR(v)}{L_v}$$

N: numero totali di pagine nel web

λ: parametro di *tuning* per favorire/sfavorire la scelta di una pagina casuale

B_u: insieme delle pagine che hanno un arco entrante in u

L_v: numero di archi uscenti da v

PAGERANK

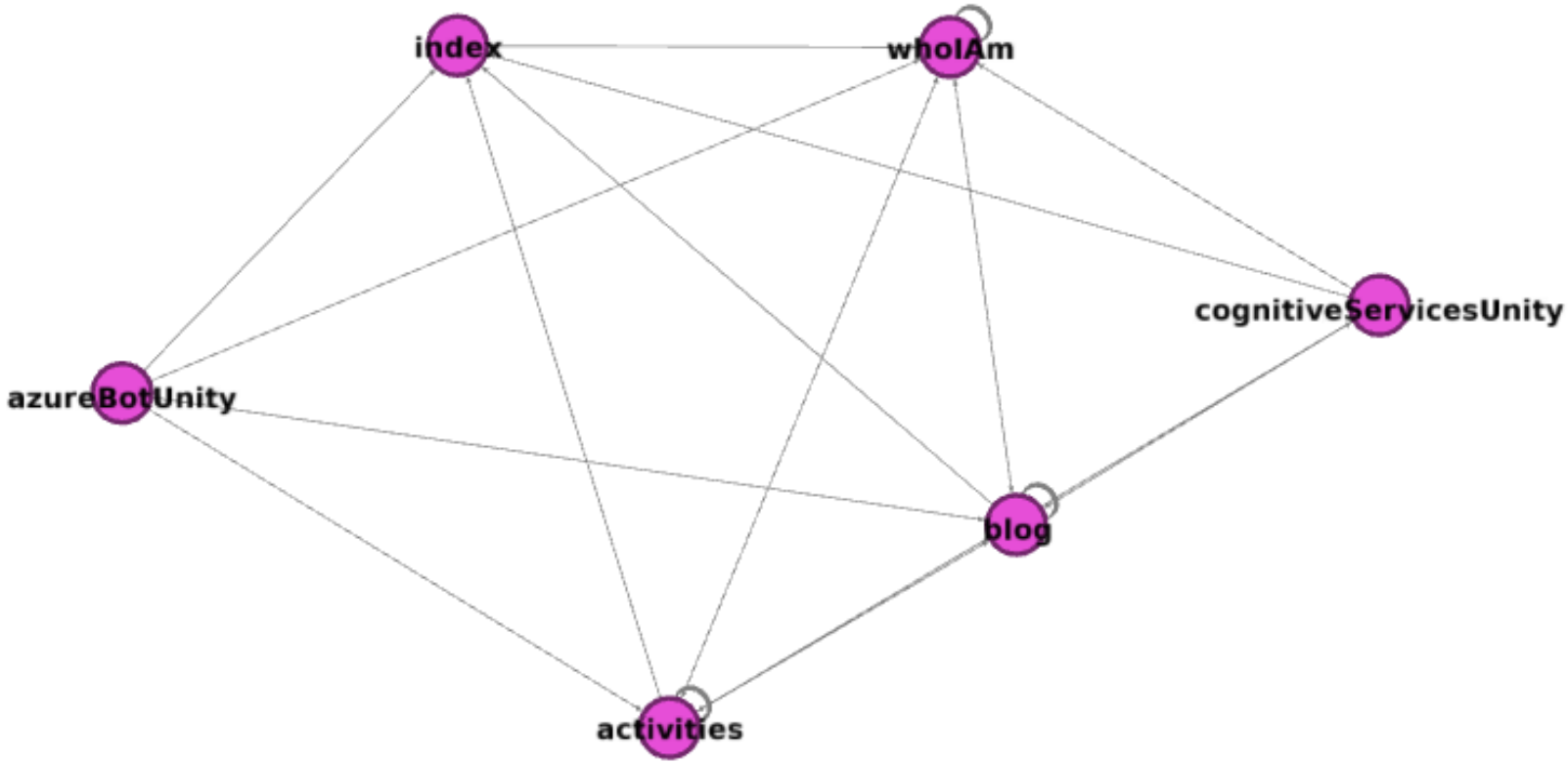
1. il random surfer si trova su una pagina *x*
2. è calcolato il pagerank di *x*
3. si sceglie un valore random *r*
   - se *r* è minore di λ ci si sposta su una pagina random
   - se *r* è maggiore di λ ci si sposta su una pagina scegliendo un link random di *x*
4. si torna al punto 1

Termino quando l'aggiornamento dei valori tra una iterazione e l'altra è molto piccolo

PAGERANK


Mario Cuomo
HOME
BLOG[IT]
ACTIVITIES
WHO I AM
Blog
Azure Bot in Unity
Come collegare semplicemente un Azure Bot in Unity, Marzo 21, 2021
VISUALIZZA L'ARTICOLO
Riguardo me
Mario Cuomo
Unity Student Ambassadors &
Microsoft Student Ambassadors

PAGERANK


index
wholAm
cognitiveServicesUnity
azureBotUnity
blog
activities


DEMO

HITS – Hyperlink Induced Topical Search

Algoritmo utilizzato per trovare due metadati – l'*authority* e l'*hub* – delle pagine web.

Authority: quanto è *autorevole* il contenuto di una pagina
Hub: quanto è autorevole il contenuto delle pagine *puntate*

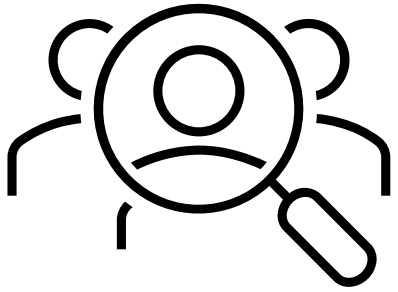
ASSUNZIONE
Buoni hub puntano a buone authority.
Buone authority sono puntate da buoni hub.

HITS

Algoritmo *agnostico* rispetto al contenuto.

Utile per diverse situazioni:

- Aggiungere metadati alle risorse
- Espandere la search
- Identificare entità esperte
- Identificare *community*



HITS

$$A(p) = \sum_{q \to p} H(q)$$

$$H(p) = \sum_{p \to q} A(q)$$

$p \to q$ indica che è presente un arco da p a q

HITS

1. $\mathrm{H}(\mathrm{q})$ e $\mathrm{A}(\mathrm{q})$ sono posti a 1 per ogni pagina q
2. per ogni pagina q si aggiorna il valore di *Authority*
3. per ogni pagina q si aggiorna il valore di *Hub*
4. si normalizzano i valori in $[0,1]$
5. si torna al punto 2

Termino quando l'aggiornamento dei valori tra una iterazione e l'altra è molto piccolo


DEMO

Microsoft Learn
Student Ambassadors

RISORSE

- *github.com/mariocuomo/talks*
- *github.com/mariocuomo/pageRank-HITS*

Grazie