

Gestione dell'Informazione

Costruzione di un thematic search engine su documenti multi-sorgente



Presentazione del progetto AA 2021/2022

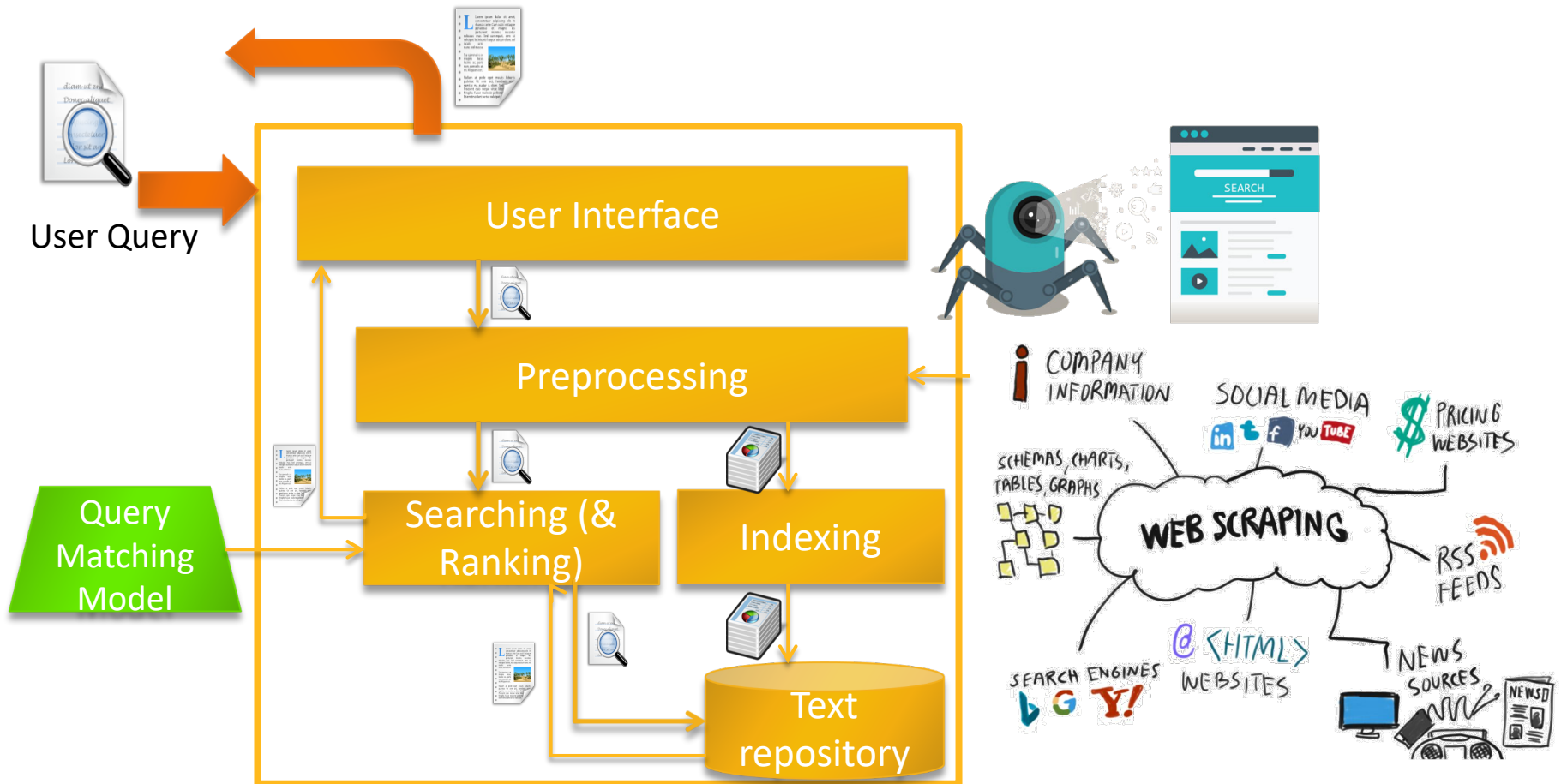
Il progetto in breve

OBIETTIVO: Sviluppare un search engine verticale su un argomento specifico di vostro interesse che si basi su un corpus di documenti provenienti da più sorgenti

COLLEZIONE DI DOCUMENTI: Costruita a partire da **almeno 2 sorgenti di documenti distinte** che trattano l'argomento scelto. Le sorgenti dati potranno essere siti web, collezioni di documenti in pdf, collezioni di documenti in formati standard W3C come XML o RDF, ... Ogni **entità del mondo reale** presente nella collezione potrà essere rappresentata in più di una sorgente

ACCESSO ALLA COLLEZIONE: L'utente avrà la possibilità formulare richieste sulla base di un linguaggio di interrogazione definito. I risultati di ogni richiesta inoltrata al sistema dovranno essere presentati in una lista ordinata in ordine decrescente di rilevanza dove ogni entità del mondo reale dovrà essere presente una e una sola volta e dovrà essere visualizzato tutte le informazioni reperite su quell'entità nelle varie sorgenti.

Architettura del sistema



Raccolta dei documenti

- ▶ Necessario individuare 2 o più sorgenti di documenti per il tema d'interesse
- ▶ La collezione dei documenti sarà costruita a partire da pagine Web, documenti in PDF, documenti in formati standard W3C come XML o RDF, ...
- ▶ La fase di pre-processing del singolo documento sarà preceduta da una fase di estrazione del contenuto testuale
- ▶ In base al formato della sorgente dati, sarà necessario usare tool diversi, ad esempio:
 - ▶ parser in grado di processare documenti nei diversi formati
 - ▶ Web API
 - ▶ Web crawler per effettuare il download di pagine web
 - ▶ Web scraper per l'estrazione di contenuti da pagine web

Preprocessing, Indexing e Query language

- ▶ I documenti di ogni sorgente dovranno essere processati e indicizzati sotto forma di text item in un index
- ▶ Ogni «**entità del mondo reale**» come ad esempio un prodotto o un articolo scientifico potrà essere rappresentata in più sorgenti e quindi potrà essere rappresentata da più di un text item
 - ▶ Ogni text item potrà fornire una visione complementare o concorrente della stessa entità del mondo reale
- ▶ Il linguaggio d'interrogazione dovrà includere almeno
 - ▶ La ricerca per keyword
 - ▶ La ricerca focalizzata sui field dello schema di indicizzazione dei vari indici

Query processing & ranking

- ▶ Data una query Q e più inverted index I_1, \dots, I_n , il query processing dovrà riscrivere Q in n query Q_1, \dots, Q_n da inoltrare ai corrispondenti indici
- ▶ Il sistema otterrà così n liste di text item che soddisfano le query ordinate in ordine decrescente di rilevanza
- ▶ La stessa «entità del mondo reale» potrà essere presente in più liste
- ▶ Il query processing dovrà prevedere un meccanismo di **identificazione della stessa entità del mondo reale** (entity resolution)
- ▶ Il query processing dovrà implementare un meccanismo «**fusione**» **delle informazioni sulla stessa entità del mondo reale** descritta da più text item
- ▶ Il query processing dovrà implementare un meccanismo di «**fusione**» **dei ranking** dei risultati restituiti dai singoli indici per ottenere un unico ranking da mostrare all'utente

Benchmarking

- ▶ Gruppo di 10 query da eseguire sul search engine
- ▶ Ogni richiesta dovrà essere descritta in linguaggio naturale e quindi tradotta nel linguaggio d'interrogazione
- ▶ Ogni richiesta avrà una caratteristica particolare in modo da mettere in evidenza le peculiarità del search engine
 - ▶ Linguaggio d'interrogazione
 - ▶ Query processing

Realizzazione e consegna del progetto

- ▶ Il progetto deve esser svolto in gruppi di preferibilmente 3 persone (o anche 2 persone)
- ▶ Al termina il gruppo dovrà produrre:
 1. un archivio (ZIP) contenente
 1. il codice realizzato
 2. Il benchmark
 3. README per l'installazione e l'uso dell'applicazione e l'esecuzione del benchmark e lettura dei risultati ottenuti eseguendo el query del benchmark
 2. una presentazione
- 1. Da consegnare una settimana prima dell'appello in cui verrà presentato il progetto
- 2. Da consegnare il giorno dell'appello

La presentazione

- ▶ Il numero di slide deve essere commisurato al tempo e comunque non superiore a 20 slide
- ▶ La presentazione deve
 1. Descrivere le sorgenti dati
 2. Descrivere lo schema degli indici
 3. descrivere il linguaggio di interrogazione
 4. Descrivere il query processing
 5. Descrivere il benchmark e le sue peculiarità
- ▶ Nel mostrare le soluzioni progettate è importante essere chiari su «**come**» il problema è stato risolto ovvero descrivere «**quale**» **soluzione tecnica** è stata individuata (approccio funzionale, metodologico) mentre non è necessario mostrare il codice, se non dei piccoli frammenti

Presentazione del progetto

- ▶ Il gruppo presenterà il progetto in occasione di un appello d'esame
 - ▶ Tempo 20 minuti per la presentazione (è molto importante rispettare i tempi)
 - ▶ Tutti i componenti del gruppo dovranno partecipare alla presentazione
- ▶ Il progetto può essere presentato in qualsiasi appello d'esame e il voto avrà validità fino a febbraio 2023
- ▶ *Non è necessario* aver superato l'esame propedeutico «Algoritmi e strutture dati» per presentare il progetto mentre è *obbligatorio* per sostenere la prova scritta

L'esame...

- ▶ 60% del voto finale dipenderà dal voto dello scritto
 - ▶ Domande aperte e semplici e brevi esercizi sugli argomenti del corso
 - ▶ 40% del voto finale dipenderà dal voto del progetto e della presentazione
 - ▶ Il voto del progetto sarà personale
 - ▶ Il voto dipenderà dalla presentazione
 - ▶ Il voto dipenderà dalla qualità e quantità del lavoro svolto .
- Aspetti valutati:
- ▶ Tipologia, ricchezza e ampiezza delle sorgenti dati
 - ▶ Progettazione e implementazione del search engine
 - ▶ Meccanismo di query processing implementato
 - ▶ Approccio alla costruzione del benchmark