Progetto Programmazione avanzata e paradigmi
Indicizzazione documenti

Luca Guerra

May 27, 2014

1 Introduzione

Si vuole sviluppare l'indicizzazione di diversi documenti di testo, in modo da sapere velocemente quali documenti contengono certe parole.

2 Visione del problema

Si possono individuare i seguenti blocchi di lavoro (Task) per il problema in esame:

- individuare i file da indicizzare
- indicizzare questi file
- cercare l'occorrenza cercata nei vari file

I primi due Task possono essere realizzati in parallelo in maniera concorrente, mentre il terzo dovrà attendere il termine dei primi due (per evitare di dare una risposta non completa).

3 Soluzione al problema

Per rendere il sistema più performante, ho deciso di sfruttare una **BlockingDeque**, questa verrà utilizzata come contenitore dei vari documenti da indicizzare. Il flusso di lavoro darà il seguente:
Allo start dell'indicizzazione avrò un processo che naviga tutto l'albero partendo dalla root (basta un unico thread, perchè il lavoro di navigare la root sarà molto più veloce del lavoro di indicizzazione) aggiungerà file alla lista, i vari processi indicizzatori rimarranno in attesa di nuovi file, appena questi saranno presenti li prenderanno e inizieranno a indicizzare in una hashtable, in questo caso questa sarà la nostra memoria condivisa, e l'accesso a questa dovrà essere mutualmente esclusivo. Per indicizzare nella hashtable avremo la parola come chiave, e come

valore una lista di stringhe che conterranno tutti nomi dei documenti che contengono la stringa specificata.
Terminata questa prima parte, il sistema è pronto a rispondere alle varie query e consultando la tabella rispondere velocemente(anche questo verrà fatto da un solo thread).

4 Conclusion

Write your conclusion here.