



UNIVERSITÀ
DEGLI STUDI
DI PALERMO



Seminari
del
Dipartimento
di
Matematica
e
Informatica

Annuncio di Seminario

MaxCover e pagine web essenziali

Paolo Boldi, Università di Milano

Mercoledì 20 aprile 2016, ore 15.00
Aula 7, Via Archirafi 34, 90123 Palermo

In questo seminario, tratteremo del rapporto di approssimazione dell'approccio greedy per la soluzione del problema MaxCover, mostrando che su istanze prese dal mondo reale esibisce un comportamento molto migliore del lower bound teorico. Come esempio guida, affronterò il problema della stima dell'indice necessario per i motori di ricerca del web per rispondere al più gran numero di query sfruttando la differenza marcata tra query e frequenze dei clic. Forniremo una possibile definizione formale per la nozione di pagine web essenziali come quelle che coprono una larga frazione di query distinte, gettando le basi del problema di trovare pagine web essenziali per una versione di MaxCover.

Nonostante in generale MaxCover sia approssimabile con un fattore $1-1/e$ ovvero circa 0.632 rispetto all'ottimale, forniremo una condizione per cui l'algoritmo greedy trova il miglior cover possibile (o rimane a un fattore limitato conosciuto da questo). Il controllo extra per l'ottimalità (o per limitare il rapporto dal valore ottimale) ha un costo algoritmico trascurabile. Inoltre, nella maggior parte delle istanze in pratica, l'algoritmo è in grado di fornire soluzioni dimostrabili essere ottimali, o vicine al valore ottimale. Collegheremo questo fenomeno al grafo dei clic delle query. I risultati sperimentali confermano che un piccolo numero di pagine web può rispondere a una larga parte delle query (ad esempio, 0.4% delle pagine può rispondere al 20% delle query).

Questo approccio può essere usato in diverse applicazioni di ricerca correlate, e ha di fatto un carattere ancora più generale – come primo esempio, il nostro studio sperimentale preliminare conferma che il nostro algoritmo ha performance molto buone su altre istanze di MaxCover (basate su social network).

Per maggiori informazioni:

Camillo Trapani

T 091 238 91062

camillo.trapani@unipa.it

Short bio: Paolo Boldi ha ottenuto il Dottorato di ricerca in informatica all'Università degli studi di Milano, dove è attualmente Professore associato. I suoi interessi di ricerca hanno riguardato diverse aree dell'informatica teorica e applicata, come ad esempio: teoria del dominio, teoria non classica della calcolabilità, calcolo distribuito, reti anonime, senso dell'orientamento, sistemi auto-stabilizzanti. Recentemente il suo lavoro si è concentrato su problemi di graph mining, information retrieval a algoritmi per i big data, campi in cui le sue ricerche hanno anche prodotto applicativi largamente utilizzati dalle persone che lavorano nello stesso campo.

Tutti gli interessati, in particolare gli studenti, sono invitati a partecipare