UNIVERSITÀ DEGLI STUDI DI PALERMO

d.m.i
matematica e informatica @ unipa

Seminari del Dipartimento di Matematica e Informatica

Seminar Announcement

MaxCover and Essential Web Pages

Paolo Boldi, Università di Milano

Wednesday, 20th April 2016, 3 p.m.
Aula 7, Via Archirafi 34, 90123 Palermo

In this talk, I will discuss the approximation ratio of the greedy approach for solving MaxCover, showing that on real-world instances it provides much better behavior than the theoretical lower bound. As a guiding example, I will address the problem of estimating the index size needed by web search engines to answer as many queries as possible by exploiting the marked difference between query and click frequencies. I provide a possible formal definition for the notion of essential web pages as those that cover a large fraction of distinct queries, casting the problem of finding essential web pages to a version of MaxCover.

Although in general MaxCover is approximable to within a factor of 1-1/e ≅ 0.632 from the optimum, I provide a condition under which the greedy algorithm does find the actual best cover (or remains at a known bounded factor from it). The extra check for optimality (or for bounding the ratio from the optimum) comes at a negligible algorithmic cost. Moreover, in most practical instances of this problem, the algorithm is able to provide solutions that are provably optimal, or close to optimal. I relate this observed phenomenon to some properties of the queries' click graph. The experimental results confirm that a small number of web pages can respond to a large fraction of the queries (e.g., 0.4% of the pages answers 20% of the queries).

This approach can be used in several related search applications, and has in fact an even more general appeal — as a first example, our preliminary experimental study confirms that our algorithm has extremely good performances on other (social network based) MaxCover instances.

Short bio: Paolo Boldi obtained his PhD in Computer Science at the University of Milano, where he is currently Associate Professor. His research interests touched many different topics in theoretical and applied computer science, such as: domain theory, non-classical computability theory, distributed computability, anonymous networks, sense of direction, self-stabilizing systems. Recently, his works focused on problems related to graph mining, information retrieval and big data algorithmics, fields where his research has also produced software tools used by many people working in the same area.

Per maggiori informazioni:
Camillo Trapani
T 091 238 91062
camillo.trapani@unipa.it

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