1. Record Nr. TD16025460

Autore	Guerrieri, Alessio
Titolo	Distributed Computing for Large-scale Graphs [Tesi di dottorato]
Editore	University of Trento, 2015-12-18
Lingua di pubblicazione	Non definito
Formato	Tesi di dottorato
Livello bibliografico	Monografia
Note	In relazione con http://eprints-phd.biblio.unitn.it/1613/

Sommario

The last decade has seen an increased attention on large-scale data analysis, caused mainly by the availability of new sources of data and the development of programming model that allowed their analysis. Since many of these sources can be modeled as graphs, many large-scale graph processing frameworks have been developed, from vertex-centric models such as pregel to more complex programming models that allow asynchronous computation, can tackle dynamism in the data and permit the usage of different amount of resources. This thesis presents theoretical and practical results in the area of distributed large- scale graph analysis by giving an overview of the entire pipeline. Data must first be pre-processed to obtain a graph, which is then partitioned into subgraphs of similar size. To analyze this graph the user must choose a system and a programming model that matches her available resources, the type of data and the class of algorithm to execute. Aside from an overview of all these different steps, this research presents three novel approaches to those steps. The first main contribution is dfep, a novel distributed partitioning algorithm that divides the edge set into similar sized partition. dfep can obtain partitions with good quality in only a few iterations. The output of dfep can then be used by etsch, a graph processing framework that uses partitions of edges as the focus of its programming model. etsch’s programming model is shown to be flexible and can easily reuse sequential classical

	graph algorithms as part of its workflow. Implementations of etsch in hadoop, spark and akka allow for a comparison of those systems and the discussion of their advantages and disadvantages. The implementation of etsch in akka is by far the fastest and is able to process billion-edges graphs faster that competitors such as gps, blogel and giraph++, while using only a few computing nodes. A final contribution is an application study of graph-centric approaches to word sense induction and disambiguation: from a large set of documents a word graph is constructed and then processed by a graph clustering algorithm, to find documents that refer to the same entities. A novel graph clustering algorithm, named tovel, uses a diffusion-based approach inspired by the cycle of water.
Localizzazioni e accesso	http://memoria.depositolegale.it/*/http://eprints-phd.biblio.unitn.it/1613/1/main.pdf