Methods, Tools and Applications for Scalable Data Analysis

LORIS BELCASTRO
DIMES, UNIVERSITY OF CALABRIA
RENDE (CS), ITALY
LBELCASTRO@DIMES.UNICAL.IT

ADVISOR: PROF. DOMENICO TALIA

Abstract

In the last years, the ability to produce and gather data has increased exponentially. In fact, thanks to the growth of social networks and the widespread diffusion of mobile phones, every day millions of people access social network services and share information about their interests and activities. Those data volumes, commonly referred as Big Data, can be exploited to extract useful information and to produce helpful knowledge for science, industry, public services and in general for humankind. However, the huge amount of data generated, the speed at which it is produced, and its heterogeneity, represent a challenge to the current storage, process and analysis capabilities. We focused on the problems of overcoming issues related to Big Data and getting valuable information and knowledge in shorter time. With regard to this, we worked on several research topics to develop and exploit novel models, methodologies, and systems for Big Data analysis, especially on Clouds, to support scalable distributed knowledge discovery applications.

As first result, we worked on the use of the MapReduce programming model for processing large datasets on Clouds, focusing on how the MapReduce paradigm can be used in combination with the workflow paradigm to enable scalable data analysis on Clouds. In particular, we worked on integrating the MapReduce paradigm in the Data Mining Cloud Framework (DMCF), a workflow framework developed at University of Calabria. In such way, DMCF's workflows can include MapReduce tasks that can be executed in parallel to support scalable data analysis on Clouds.

We also worked on methodologies for extracting knowledge from data gathered from social networks, with particular regards to the implementation of a novel data mining technique for extracting Regions-of-Interest (RoIs) from geotagged social data with a high accuracy. The analysis of geo-tagged data allows to determine if users visited or not interesting locations (e.g., touristic attractions, shopping malls, squares, parks), often called Places-of-Interest (PoIs). Since a PoI is generally identified by the geographical coordinates of a single point, it is hard to match it with user trajectories. For this reason, it is useful to define the RoI representing the boundaries of the PoI's area. With regard to this, we proposed a novel RoI mining technique, called G-RoI, which differs from the existing approaches as it exploits the indications contained in geotagged social media items (e.g. tweets or posts with geospatial information) to discover the RoI of a PoI. Many experiments have been performed to assess the accuracy of G-RoI over real geotagged items extracted from Flickr. The experimental results show that G-RoI is more accurate in identifying RoIs than existing techniques.