Heterogeneous social data management and mining

CDS 2.0

Nacéra BENNACER, Francesca BUGIOTTI, Gianluca QUERCINI

LRI Research Laboratory

nacera.bennacer@lri.fr, francesca.bugiotti@lri.fr, gianluca.quercini@lri.fr

1 Context

Online social networks, such as Facebook, Twitter, Linkedin, and Weibo, are widely-used platforms that enable key social and professional interactions among millions of users across the world. These data are largely noisy, heterogeneous (e.g., graph data and textual content), uncertain (e.g., the opinions of users), sometimes false (e.g., rumors or intentional lies), unstructured and incomplete. Also, these data are multilingual by nature and often ambiguous; this is the case, for example, of the toponyms, such as "Paris", that could refer to different actual locations across the world or names of people. Our main challenges are 1) to integrate, disambiguate, and reconcile such data to extract information, by exploiting different rich Web data sources such as Linked Open Data, large encyclopaedias (e.g., Wikipedia, BabelNet) and knowledge bases (e.g., Yago, DBpedia, WordNet), and 2) to define efficient crawling, querying and storage methods by exploiting NoSQL stores and distributed environments. These challenges are important to achieve different goals, such as 1) digital identity discovery for human resources management, 2) automatic integration of the individual's profiles across different social networks to have an holistic view of the information about the individual, 3) developing new algorithms for recommendation systems based on social Web data, and 4) analyzing the correlation between social data available on an individual to his personality traits.

2 Reconciliation of profiles across Social Networks

A first objective of our research is investigating efficient methods to detect profiles referring to the same individuals across multiple social networks in order to integrate all the information regarding the individuals themselves. In a previous work, we evaluated an approach that used a set of rules to match (or, reconcile) profiles based on a limited number of information disclosed in the profiles, such as the nickname, the name and the geographic location. One key problem of this approach is that it trusts the information found in the profiles and makes no attempt to verify their correctness and thus detect false/outdated information. In the context of a research Master internship, our goal is to propose and implement an approach for the reconciliation of profiles that is both scalable and robust to false information.

3 Define efficient querying and storage methods in Cloud Environments

In the context of integrating data coming from multiple social networks it is important to define a methodology that allows us to efficient store and query data. In this direction we want to study if and how it is possible to store and query data in an heterogeneous environment multi-cloud. The study will consider the characteristics of different Cloud Service Platforms such as Amazon Web Service and Microsoft Azure and will try to analyze how and when it is better to store data in one or the other service with respect to the desired queries. A possible analysis can start from the consideration that Amazon Dynamo DB does not allow the definition of indexes. The first objective in this direction is to study for our scenario if it is more convenient to organize data in an ad-hoc way in Amazon Dynamo DB (adding structures that simulate indexes if necessary), if it is better to store our data in Microsoft Azure (adding also in this case the necessary support structures), or to use multiple platforms and smartly redirect queries to the various systems.