Advanced Information Systems
project status report

Gasparella Luca

A.y. 2008/2009

Problem definition

A soaring number of services bases their functionality on a distribuited context where each local implementation is often realized separately from the others and the model used to represent data may be significatively different. In particular the problem analyzed take into the syntactic heterogeneity that fold several aspects about the definitions for representing data.

The main aim of the project is to allow the execution of queries on datasources which have been described differently through two distinct schemas.

Status

The syntactic heterogeneity is a part of the problem known as "Schema matching" which objective is finding a semantic relation between two objects. In order to solve the problem, a solution based on "stacking" from machine learning was adopted in [3] achieving a very good results since the accuracy of their proposal is around 71-92%. They average a learning weight based on the attribute and its possible values.

The problem that comes directly is the effective syntactic difference between some field that may be, for example, the definition of the concept city[1] where the following are equivalent: "city","coordinates[2]" and "postal code[3]". In order to best fit the tasks it would be considered to adopt a graph-based (not oriented) structure where each node represents an attribute and an arc connecting another node indicates the possible association between them; the weight of the arc corresponds to a probabilistic indicator of similiarity obtained usind the "stacking" method indicated above and the results coming from [1] adopting the token-based metrics technique from [2]. The primary target will be to identify from the data the meaning of the field name.

To support the construction of graph-based structure may be used several internet site that offers free services like *Wordnet*[4] and *Wikipedia*[5] since are considered good enough to the project's scope of retrieving similiarities about a certain word.

Currently the idea is to define a common meta-structure which can be easily mapped to the sources structures that will be build as product of the graph-structure chosing the best "accuracy" value. Next step is to define the correct approach in order to achieve a efficient classification method that may correlate two simple schemas for which data are known.

[1] A right and proper assumption concerns the language used to define the model and for simplicity is english.

[2] The geographic coordinate system.

[3] In the case thought is italian, but the concept is similar to Americna's ZIP code.

References

[1] Xin Dong, Alon Halevy, *"Indexing Dataspaces"*, SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data (2007), pp. 43-54.

[2] AK Elmagarmid, PG Ipeirotis, VS Verykios, *"Duplicate Record Detection: A Survey"*, Knowledge and Data Engineering, IEEE Transactions on, Vol. 19, No. 1. (2007), pp. 1-16.

[3] AnHai Doan, Pedro Domingos, Alon Halevy, *"Learning to Match the Schemas of Data Sources: A Multistrategy Approach"*, Kluwer Academic Publishers, Volume 50 , Issue 3 (2003), pp. 279 - 301.

[4] *http://wordnetweb.princeton.edu*, 31-05-2009.

[5] *http://wikipedia.org*, 31-05-2009.