

An approach for matching schemes of heterogeneous relational databases.

ABSTRACT

Schema matching is a basic problem in many database application domains, such as data integration. The problem of schema matching can be formulated as follows, "given two schemas, S_i and S_j , find the most plausible correspondences between the elements of S_i and S_j , exploiting all available information, such as the schemas, instance data, and auxiliary sources". Given the rapidly increasing number of data sources to integrate and due to database heterogeneities, manually identifying schema matches is a tedious, time consuming, error-prone, and therefore expensive process. As systems become able to handle more complex databases and applications, their schemas become large, further increasing the number of matches to be performed. Thus, automating this process, which attempts to achieve faster and less labor-intensive, has been one of the main tasks in data integration. However, it is not possible to determine fully automatically the different correspondences between schemas, primarily because of the differing and often not explicated or documented semantics of the schemas. Several solutions in solving the issues of schema matching have been proposed. Nevertheless, these solutions are still limited, as they do not explore most of the available information related to schemas and thus affect the result of integration. This paper presents an approach for matching schemas of heterogeneous relational databases that utilizes most of the information related to schemas, which indirectly explores the implicit semantics of the schemas, that further improves the results of the integration.

Keyword: Biomedical database; Database integration; Heterogeneous; Schema matching.