

Overlap Discovery among Heterogeneous Databases Schemata using Link-based Analysis

Leena Al-Hussaini - joint work with: Malcolm Atkinson, Dave Berry, and Stratis Viglas

leena@nesc.ac.uk



Introduction

We study the "structure" of a database schema to shed light on important aspects in an individual database as well as when a number of database schemas are to be integrated.

Past studies have focused on either integrating single relations, or have applied transformation rules to convert a specific database schema to another. The latter forces one database schema to be adapted to another database schema, which can sometimes be rendered inefficient or not feasible due to the complexity of the database. To integrate database schemas of multiple relations, we follow a different approach.

Our approach is based on inferring relations that are likely to overlap from two or more database schemas by investigating the way similar relations in different databases are linked. We hope this leads to a way of efficiently integrating database schemas in a scalable manner.

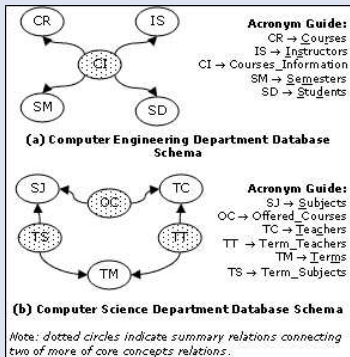


Figure 1: Various Designs of Databases Schemas

To motivate the problem, Fig. 1 sketches a pair of simple database schemas. The scenario in Fig. 1 presents the database schemas of two university departments: Computer Engineering department (a)

and Computer Science department (b). After many years of successful collaborative research, both departments have decided to merge into one department: Computer Science and Engineering.

By way of example, some interesting challenges that arise are: *How can one discover that Courses (CR) in Fig. 1 (a) should be integrated with Subjects (SJ) in Fig. 1 (b)? How can one discover that Courses_Information (CI) in Fig. 1 (a) overlaps with relations Offered_Courses (OC), Term_Teachers (TT), and Term_Subjects (TS) in Fig. 1 (b)?*

We investigate two approaches.

Approach 1: Classification and Overlap Discovery

In this approach, we break our study of a database structure into two phases:

We first distil what we term core concept relations from summary relations. This is achieved by using the HITS ranking algorithm [Kleinberg, ACM Journal, 99]. HITS classifies pages on the Web as authoritative sources and hub sources based on a mutually reinforcing relationship. That is, a good authority page is usually linked to by a set of good hub authorities. We use this algorithm to classify core concept relations and summary relations as depicted in Fig. 2. Indeed, we regard core concept relations as authorities and summary relations as hubs. So, Fig. 2 consists of the classification result of Fig. 1 (a) on the left-hand side of Fig. 2 and Fig. 1 (b) on the right-hand side of Fig. 2.

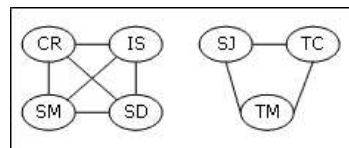


Figure 2: Relations Classification

We then operate only on authority sources and build an authority database graph (ADG) for each database. We attach informative weights on both the nodes and the edges of the ADG, based on the ranking obtained during the first phase. Then, we walk each graph and build a transition matrix from where we try to infer overlaps among two or more database schemas. Fig. 3 presents the result of walking the two ADGs from Fig. 2 and how overlapping authorities can be discovered.

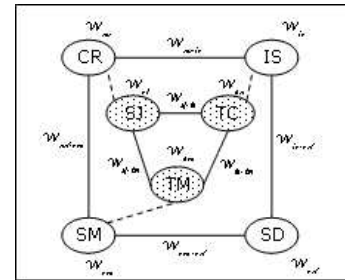


Figure 3: Overlap Discovery

Approach 2: Overlap Discovery without Classification

In this approach, we investigate the structure of a database without distinguishing between the different types of relations that forms it. We try to discover which relation(s) in a database is/are likely to match with which other relation(s) in another database. For this approach, we extend Google's PageRank [Brin et al, Computer Networks, 98] to work with relational database schemas; in an algorithm we call RelationRank. The basic PageRank algorithm is defined as follows:

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

Briefly describing the algorithm: A is a web page (in our setting, A is a relation) linking to pages (relations) T_1 to T_n , d is a damping factor, and $C(T_1)$ is the number of links referring to T_1 .

Fig. 4 illustrates the schemas of the two departments and what set of relations from the schema of one department overlaps with the schema of another department. Each node in both departments graphs is attached with rank value generated by using the proposed RelationRank algorithm. For our purposes, we extend the original PageRank algorithm to measure the number of attributes of each relation. This produces unique rank values associated with each relation. Rank values are normalized in each graph. Euclidean distance is used as the metric to find the minimum distance between two or more of overlapping relations.

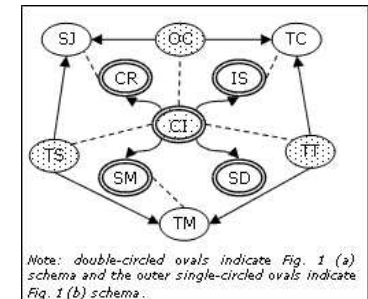


Figure 4: Overlap Discovery without Classification

Conclusion

We have proposed novel approaches to discovering overlap among heterogeneous database schemas. The approaches are based on link-based analysis and extend well-known ranking algorithms. We hope these ideas lead to efficient and scalable algorithms for data integration.