SCHEMA BASED DEDUPLICATION

(Research-in-Progress)

Pei Li, Andrea Maurino
University of Milan, Bicocca
pei.li,maurino@disco.unimib.it

Abstract: In this paper we present a preliminary report on a domain independent strategy to reduce duplicated records by means of the knowledge stored in the schema. According to different kinds of relationships, we propose specific techniques to build and compare the knowledge networks by means of graph-based similarity techniques.

1. INTRODUCTION

Deduplication is the activity of detecting if different records of the same dataset represent the same real world entity. In a deduplication process, candidate record pairs usually fall into three sets: i) records that definitively match, ii) records that definitively do not match, and iii) records that possibly match.

In this paper, we focus on the definition of a domain independent strategy [1] for reducing the possible match record sets based on the use of knowledge stored in the whole schema and not only restricted to the analyzed entity. Correspondingly, 1) determining what kinds of relationships could be taken into account, 2) how it is exploited to form a knowledge network of each candidate record, and 3) which similarity measures should be applied are the main focuses in this paper.

2. SCHEMA-AWARE DEDUPLICATION

Records in a database are never alone but always belong to certain knowledge networks (KN) consisting of a group of records connected with each other via entity relationships. When a new record is inserted to the database, due to the scattered information and dirty data issues related to an existing object, a part of its KN is newly created, while the rest remains an existing part of the object's KN it represents. From this point of view, the similarity of two records relies on the similarity of their KNs. We first give a formal definition of knowledge network:

DEFINITION 1 (Knowledge Network). A knowledge network (KN) of a tuple in database consists of the possible match tuple that can be duplicated and a group of associated records connected to it via relationships stored in the database. It is denoted by a directed graph $G =< V, E >$, where nodes $V = \{v_i\}_{i=1}^{n}$ encode the involved records and the generic edge $e_{ij} \in E$ represents a certain instance of the relationship from v_i to v_j . $|V|$ is the number of nodes in the graph.

Let G_a be the knowledge network graph created from the record a, G_b a KN created from record b, and $s(G_a, G_b) \in[0..1]$ is a similarity function between the two KNs, where $s(G_a, G_b) = 1$ if a = b. In the following subsections, we present different strategies for the most adopted types of relationships and we present how it is possible to build the KNs taking into account schema cardinalities and corresponding similarity metrics for $s(G_a, G_b)$. It is worthwhile to notice that in real-world situations combinations of them can be applied to deduplication tasks where the underlying schema has complex structures.

2.1 Many to Many Relationship

KNs of records exploiting the (0/) 1: n-(0/) 1: m relationship consists of associated records that are linked by iteratively expanded relationship paths. With the lengths of such schema paths increasing, the involved records are considered as less important to the KN. The similarity of two KNs, is measured by their common subgraphs. We apply the similarity metric introduced in [2] to compare two graphs.

2.2 Many to One Relationship

In the case of (0/)1 : n-1 : 1 or 1:1 – (0/)1:n relationship, we propose to measure the similarity between records by means of SimRank [3]. This similarity can be calculated by putting together KNs of the two candidates to form a bipartite graph, where nodes are merged if their values are semantically the same.

2.3 One to One Relationship

KNs of records can be extended in attribute level if having (0/) 1:1-(0/) 1:1 entity relationships with other records. In this case it is possible to merge the two tables and to perform traditional pairwise analysis.

2.4 A Methodology for Schem-Based Deduplication

In this section we present a methodology supporting the use of the above mentioned techniques, which is composed by the following phases:

1) Identify the target entity.
2) Select all relationships related to the target entity.
3) For each identified relationship assign a weight w_i so that the sum of all weights is equal to 1.
4) Check if it is possible to augment the target entity with other ones connected by a relationship with a (0/1):1 – (0:1):1 cardinality (see section 2.3).
5) Perform FBS analysis over the augmented target entity.
6) For each pair of possible match records and for each identified relationship apply a similarity metric according to the cardinality of relationship, resulting in a similarity score s_i.
7) Combine s_i to calculate an overall similarity score $s = \sum w_i s_i$.
8) Apply threshold values to check if two records represent the same object.

3. CONCLUSIONS

In this paper we report the preliminary results of a technique for reducing possible match sets by means of schema aware analysis, using knowledge networks and graph similarity measures. Our further work includes a complete analysis of relationships and evaluation of computing complexity of KNs.

REFERENCES

[1]. Pei Li, Andrea Maurino, Schema-based deduplication, QDB '09

[2]. S. Q. Le, et al. A novel graph-based similarity measure for 2d chemical structures. Genoome Informatics, 15(2), 2004.

[3]. G. Jeh, J. Widom. Simrank: a measure of structural-context similarity. KDD '02, pages 538–543.