Duplicate Detection of Query results from Multiple Web Databases

Perspectives

Article 6
Year of Publication: 2011

Authors:
Hemalatha S
Raja K
Tholkappia Arasu

Abstract

The results from multiple databases compose the deep or hidden Web, which is estimated to contain a much larger amount of high quality, usually structured information and to have a faster growth rate than the static Web. The system that helps users integrate and, more importantly, compare the query results returned from multiple Web databases, an important task is to match the different sources’ records that refer to the same real-world entity. Most state-of-the-art record matching methods are supervised, which requires the user to provide training data. In the Web databases, the records are not available in hand as they are query-
dependent, they can be obtained only after the user submits the query. After removal of the same-source duplicates, the assumed non duplicate records from the same source can be used as training examples. The method uses the classifiers the weighted component similarity summing classifier (WCSS) and Support Vector Machine (SVM) classifier that works along with the Gaussian mixture model (GMM) to iteratively to identify the duplicates. The classifiers work cooperatively to identify the duplicate records. The complete GMM is parameterized by the mean vectors, covariance matrices and mixture weights from all the records.

Reference

- W.Su, J. Wang, and Frederick H. Lochovsky, “Record Matching over Query Results from Multiple Web Databases” IEEE Transactions on knowledge and data engineering.
- Gaussian Mixture Models Douglas Reynolds MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA

**Index Terms**

Computer Science  
Databases

**Key words**

Gaussian Mixture Model  
supervised learning  
Support Vector Machine