عنوان مقاله:

Identifying Duplicate Records by Using Estimation of Distribution Algorithms to Learn the Semantics

محل انتشار:

یازدهمین کنفرانس سالانه انجمن کامپیوتر ایران (سال: 1384)

تعداد صفحات اصل مقاله: 8

نویسندگان:

Saied Haidarian Shahri[1] - *Control and Intelligent Processing Center of Excellence (CIPCE)Department of Computer and Electrical Engineering University of Tehran, Tehran, Iran*

Caro Lucas - *Control and Intelligent Processing Center of Excellence (CIPCE)Department of Computer and Electrical Engineering University of Tehran, Tehran, Iran*

Babak N. Araabi[1,2] - *School of Cognitive Sciences, Institute for studies in theoretical Physics and Mathematics, Tehran, Iran*

خلاصه مقاله:

When data is gathered from various sources to be included in integrated information systems, for example data warehouses, the likelihood of existence of duplicate and inconsistent data records increases. A flexible and automatic reasoning mechanism is required to clean the data, to enable the user to draw accurate statistics and reports from this wealth of data, which are to be used in the decision making of entrepreneurial enterprises. In this paper, we have employed an approach for deduplication, which takes advantage of a fuzzy logic framework. The fuzzy inference system is then optimized by means of the Bayesian Optimization Algorithm, a class of Estimation of Distribution Algorithms, which can learn complex multivariate relations of bounded order. This class of algorithms is inspired form the breeder genetic algorithm, which is used in the science of livestock breeding. The experiments reveal that this approach is capable of eliminating duplicates abound with uncertainty, and therefore the resultant data is of better quality.

کلمات کلیدی:

Duplicate Elimination, Estimation of Distribution Algorithms,Fuzzy Inference System