Enhanced normalization approach addressing stop-word complexity in compound-word schema labels

ABSTRACT

An extensive review of the existing schema matching approaches discovered an area of improvement in the field of semantic schema matching. Normalization and lexical annotation methods using WordNet have been somewhat successful in general cases. However, in the presence of stop-words these approaches result in poor accuracy. Stop-words have previously been ignored in most studies resulting in false negative conclusions. This paper proposes NORMSTOP (NORMalizer of schemata having STOP-words) as an improved schema normalization approach that addresses the complexity of stop-words (e.g. 'by', 'at', 'and,' or') in Compound Word (CW) schema labels. Using a combined set of WordNet features, NORMSTOP isolates these labels during the preprocessing stage and resets the base-form to a relevant WordNet term, or an annotable compound noun. When tested on the same real dataset used in the earlier approach - (NORMS or NORMalizer of Schemata), NORMSTOP shows up to 13% improvement in annotation recall measurement. This level of improvement takes the overall schema matching process another step closer to perfect accuracy; while its absence exposes a gap in expectation, especially in today's databases, where stop-words are in abundance. –ceramic exhibit a huge potential to act as a green phosphor in opto-electronic devices.

Keyword: Database integration; Schema matching; Data heterogeneity; Semantic schema matching; Schema label normalization; Stop-words