# Efficient Approximate Dictionary Matching

Saurabh Kishore
skishore@gmail.com

Ashish V. Tendulkar
ashishvt@gmail.com

## ABSTRACT

Named entity recognition (NER) systems are important for extracting useful information from unstructured data sources. It is known that large domain dictionaries help in improving extraction performance of NER. Unstructured text usually contains entity mentions that are different from their standard dictionary form. Approximate matching is important to identify the correct dictionary entity for such variants. This is a challenging problem, as every entity in the dictionary is a candidate match for the variant. In this paper, we propose a novel approach for efficient approximate dictionary matching. The key idea is to compare a given query only against a set of most likely candidate matches from the dictionary so as to achieve substantial reduction in the number of matching operations. In order to enable this, the proposed approach first performs clustering of similar entities and then represents each cluster with a profile matrix, which stores the probability of an occurrence of a particular character at a specific location in the entity string. Thus, the dictionary is represented with a set of profile matrices, which are much smaller than the actual number of entities. A given query entity is first matched against the profiles and the clusters corresponding to top-K best scoring profiles are selected to obtain a list of most likely matching candidates. The query is then compared with each candidate match entity and the approximate match is declared if both the query and the candidate entity are within acceptable edit distance threshold. We have performed rigorous evaluation of our approach on several publicly available datasets. The proposed algorithm outperforms alternative approaches in detecting approximately matching entities for a given query using far lesser number of comparison operations.


*The 18th International Conference on Management of Data (COMAD),* 14th-16th Dec, 2012 at Pune, India.