Precision Estimation for Matching Place Names to a Digital Gazetteer

Martin Doerr[1] and Manos Papagelis[1, 2]
[1] Institute of Computer Science, Foundation for Research and Technology - Hellas
P.O. Box 1385, GR-71110, Heraklion, Greece
{martin, papaggel}@ics.forth.gr
[2] Department of Computer Science, University of Crete
P.O. Box 2208, GR-71409, Heraklion, Greece

ABSTRACT

Most of the information available digitally has some reference to locations or objects in geographic space. Digital gazetteers are commonly employed to identify the referred place names, in information integration procedures, by matching gazetteer records with "local choice of terms" that come from diverse knowledge domains. However, identification process is not always possible because gazetteers provide incomplete information compared to the world structure. This process suffers from failures due to (i) non-existence (incompleteness of a gazetteer), (ii) duplication (a geographic name is matched to more than one place), or (iii) false positive matches (the geographic place found is not the one initially intended by the place name described in the local data source; this case results in semantic inconsistency).

The previous paragraph, describes a communication problem via a shared authority, common in information integration, dissemination and sharing processes. We study the cases of success and the cases for failure of this mapping process. We are especially interested in estimating the precision of one-to-one mappings, the case in which one place name matches exactly one place, because consequently, human intervention in the matching process can be systematically reduced to the rest cases. Based on this study, we present a statistical method that permits to estimate (i) the completeness of a gazetteer with respect to a target area, (ii) the expected precision and recall of one-to-one mappings of source place names to the gazetteer (iii) the semantic inconsistency that exists in one-to-one mappings and, (iv) the degree to which, precision and recall are modified, under knowledge of the identity of higher levels in a hierarchy of places. Our methodology is based solely on statistical analysis of the results of the matching process itself, of all the place names of a data source against a gazetteer and as such, it requires no additional information and effort.

In our experimental evaluation, we employ a set of 1000 place names originating from the LUPA database, a large data source that describes all known archaeological findings of a specific kind of a large geographical target area, statistically well distributed from small villages to major cities. The third-party authority employed is the well-known Alexandria Digital Gazetteer.

As far as our knowledge of literature permits, our work is the first that provides estimations of completeness and correctness of a digital gazetteer and its significance is summarized as:

- Provides decision support for gazetteer use and gazetteer development issues
- Determines the error propagation introduced by mismatch into statistical analysis based on the matched data, such as “how many objects where found in this geographic area?”
- Defines the degree to which knowledge of the identity of higher levels in a hierarchy of places improves the automatic mapping process

Since our methodology is general, it can probably be easily adopted to matching terms against of other hierarchies of concepts.