Abstract: Applying grammatical inference techniques to activity-diary data

The aim of this paper is to illustrate the application of grammatical inference techniques in the context of space-time activity-diary data. To this end the sequential information which is embedded in the data, is used.

Grammar Induction, also known as Grammatical Inference, is a theory which has been developed by computational linguists for modelling strings of symbols. It is an instance of Inductive Learning, which can be formulated as the task of discovering common structures in examples which are supposed to be generated by the same process. Grammars were originally developed to model natural languages and have been extensively used since then in the analysis and design of computer languages and compilers. More recently, grammars have found their application in the analysis of biological sequences. A lot of fruitful results were generated in both domains, which can partially be attributed to the fact that these domains seem particularly well suited to be captured and represented in terms of formal languages.

A formal grammar within the domain of activity-based research consists of an alphabet of activities; called the terminal symbols, a second alphabet of variables; called the non-terminal symbols and a set of production rules. By consecutively applying different production rules of the grammar, activity sequences can be generated. The motivation for building this set of production rules is to make a formal model of human activities which are typically carried out.

First, a general introduction is given in the paper to the theory of formal grammars, to their properties and to their position in the Chomsky hierarchy. Hereafter, parse trees are introduced, which reflect the syntactic structure of the activity sequences. Furthermore, a distinction is made in the paper between deterministic and stochastic grammars. Deterministic grammars are described first, which purpose is to set the stage for applying stochastic grammars to the activity-diary data. Stochastic grammars are obtained by superimposing a probability structure on the production rules. Stochastic grammars appear to be particularly useful to model activity sequences since it gets increasingly difficult to create one particular pattern as more activity sequences are determined and the family grows larger. Therefore, by applying stochastic grammars, exceptions to the rules of the pattern are allowed but instead of considering all probabilities equal, less score is given to these exceptions than a strong match. The presented approach has been tested on data that were collected in 1997 in the municipalities of Hendrik-Ido-Ambacht and Zwijndrecht in the Netherlands for the Albatross model system.