Summary: In a field of research about general reasoning mechanisms, it is essential to have appropriate benchmarks. Ideally, the benchmarks should reflect possible applications of the developed technology. In AI Planning, researchers more and more tend to draw their testing examples from the benchmark collections used in the International Planning Competition (IPC). In the organization of (the deterministic part of) the fourth IPC, IPC-4, the authors therefore invested significant effort to create a useful set of benchmarks. They come from five different (potential) real-world applications of planning: airport ground traffic control, oil derivative transportation in pipeline networks, model-checking safety properties, power supply restoration, and UMTS call setup. Adapting and preparing such an application for use as a benchmark in the IPC involves, at the time, inevitable (often drastic) simplifications, as well as careful choice between, and engineering of, domain encodings. For the first time in the IPC, we used compilations to formulate complex domain features in simple languages such as STRIPS, rather than just dropping the more interesting problem constraints in the simpler language subsets. The article explains and discusses the five application domains and their adaptation to form the PDDL test suites used in IPC-4. We summarize known theoretical results on structural properties of the domains, regarding their computational complexity and provable properties of their topology under the h+ function (an idealized version of the relaxed plan heuristic). We present new (empirical) results illuminating properties such as the quality of the most wide-spread heuristic functions (planning graph, serial planning graph, and relaxed plan), the growth of propositional representations over instance size, and the number of actions available to achieve each fact; we discuss these data in conjunction with the best results achieved by the different kinds of planners participating in IPC-4.

Keywords: compilations; STRIPS