A quantitative analysis of the robustness of Knowledge-Based Systems through degradation studies

Perry Groot Frank van Harmelen Annette ten Teije

Div. of Math. and CS., Faculty of Sciences, Vrije Universiteit Amsterdam

Full paper will appear in the Knowledge and Information Systems journal.

Summary: The difference between Knowledge Based Systems (KBSs) and 'conventional' software is often claimed to be the ability to deal with incomplete, incorrect, and uncertain knowledge and data. Although nowadays this distinction is not considered to be sufficient or necessary to define a KBS, it is believed that its still an essential dimension of KBS valiadation. In this paper, we argue for the need for *quantitative analysis* of the quality of KBSs. In particular, we show how *robust behavior* in the presence of incomplete system-input as well as an incomplete and incorrect knowledge base is amenable to such quantitative analysis. Our quantitative analysis is based on the idea of *degradation studies*: analyze how the quality of the output changes as a function of degrading input. We propose a set of *general definitions* which are general enough that they can be used in similar degradation experiments by others. We show the practicality of our approach by applying it to a particular *case study* thereby yielding surpising insights into the behavior of the system under study.

Approach: The approach to quantifying the robustness of a KBS is based on a *degradation study*: gradually decrease the quality of the KBS input, and measure how the KBS output quality changes as a result. We consider the approach to be the central contribution of the paper. Of course, we must be more precise about the notion of 'quality' of the KBS input and output. We expect that each task-type will come with its own measure for input quality (for KB and data) and therefore leave it open. For the output of the KBS we assume that it can be interpreted as a *set* of answers. For many typical KBS tasks, this is a realistic assumption. The output quality is then measured using the measures *recall* and *precision*. The recall is the fraction of correct answers that the system actually computes whereas the precision is the fraction of computed answers that are actually correct. These measures are well known from the literature on information retrieval. One can also consider these measures to be gradual versions of soundness and completeness. Some balance usually has to be found between those two measure as in practice increasing one usually means the other will decrease.

Experiments: For the case study we use a KBS that classifies commonly occurring vegetation in Sourthern Germany. Given a number of observables (e.g., color of flower, size of leaves, shape of leaves, etc.) the system returns one or more plants that closely match the observables. Note that it only serves to illustrate our proposal to analyze the robustness of KBSs through degradation studies. The important aspects of this case study are the quantities we measure and how we analyze them, not the robustness results we obtain for the specific KBS we use. Underlying our experiments lies the idea that a KBS produces an output through some algorithm that accepts some input. The latter conists of data and knowledge. For the degrading input quality we look at incomplete data input (i.e., missing observations) and an incomplete and incorrect knowledge base. Furthermore, we look at the order in which data or knowledge is changed or removed. Some examples of interesting insights we found are the following:

Data input: Using the ordering of observations obtained from the test cases we obtained that the first 6 observations did not contribute to any quality increase for either the recall or precision. Interrupting the system before is has processed 6 is therefore useless. Thereafter, for about 6 to 15 observations, both measures increased rapidly and remained stable for the remaining observations. This was surprising as most cases contain 19-30 observations whereas our experiments indicate that processing more than 15 observations is not useful as it will not increase the quality of the output of the system. Using other orderings for the observations influenced our robustnes results, however, the ordering found in the test cases and the one used by the interface were surprisingly effective when compared to a random ordering.

Knowledge Base: Changing *every* rule slightly resulted in a decrease in precision of only 0.1 (on the quality interval [0,1]). Hence, the KBS seems to be robust for small errors in its knowledge base. In other experiments we looked at the robustness of the system with respect to incompleteness by removing rules from its knowledge base. When 40% of the knowledge base was removed at random there was almost no quality loss. However, when rules were removed in a biased way (figure 1) the quality of the output already started dropping after removing 15% of the rules indicating that some parts of the knowledge base had more influence on the outcome of the system.

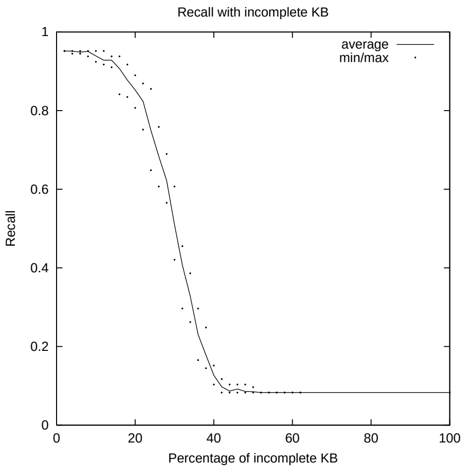

Figure 1: Recall with incomplete KB where rules were removed using a probability measure.

Final words: The ultimate suggestion that follows from this work is that any KBS should upon delivery come accompanied with a set of degradation statistics such as discussed in this paper as a quantitative way of measuring interesting and important aspects of the systems quality. this would contribute to a more empirical and quantitative analysis of AI systems and of KBSs in particular.