Lexical Homogeneity of A Rule Base

Ook Lee
College of Information and Communications
Hanyang University
Seoul, Korea
Tel: +822-2290-1087, Fax: +822-2290-1886
e-mail: ooklee@hanyang.ac.kr

Abstract: In this paper, I propose a measure of the status of a rule base that can be used to predict the degree of difficulty in the maintenance of a rule base.

1. Introduction

The content of a knowledge base makes the maintenance of Knowledge-Based Systems (KBS) seem more difficult compared to conventional software maintenance. I developed a measure of the status of a rule base that can be used to predict the degree of difficulty in the maintenance of a rule base. I tested three real-world rule bases with this measure and was able to predict different degrees of difficulty when it comes to maintain each rule base.

2. KBS Maintenance

Coenen and Bench-Capon [1] categorized KBS maintenance as shown in Table 2-1.

Table 2-1 KBS Maintenance Categories

Maintenance Category	Explanation
1. **Corrective maintenance**	Corrective maintenance of KBS refers to maintenance required because *a KBS is not* behaving as *it* should, e.g., a wrong conclusion may be drawn due to errors in encoding knowledge into the knowledge base.
2. **Adaptive maintenance**	Adaptive maintenance of KBS results from changes in the environment in which a system is designed to operate such as changes in domain knowledge.

3. **Perfective maintenance**	Perfective maintenance of KBS results from changes in user requirements such as changes in user interface.

In all three categories, KBS maintenance mostly involves updating the knowledge base. Especially for the KBSs developed from KBS shells, the only change that anybody can make really is the knowledge in knowledge bases or rule bases. Thus the concern in this paper is with changes made in the knowledge base of a KBS. In other words, maintenance of a KBS is really about maintenance of a knowledge base and it is maintenance of a knowledge base that should be distinguished from the area of conventional software.

3. Lexical Homogeneity of a Rule Base

This section explores what makes the amount of difficulty in maintaining a rule base different from one rule base to another. To answer this question, I developed a concept called Lexical Homogeneity of a rule base as the factor that predicts the level of difficulty in the maintenance of a rule base.

< Lexical Homogeneity of a rule base >

Lexical Homogeneity of a rule base is a measure of the repetitive use of terms in a rule base. A rule base can use the same terms many times in different rules. Some rule bases have the tendency to use terms repetitively while others do not. The Lexical Homogeneity of a rule base is defined as follows:

Rule Base Homogeneity= S/T

where

S=Total # of shared terms and N=Total # of terms. This definition measures the amount of repetitive use of the same terms in a rule base. Since this definition requires rather tedious computation, I devised an approximation that is simpler to compute. I now present a detailed explanation of this approximation. The lexical distance between any two rules(denoted as i, j) is defined as follows.

D[i,j]=(1-(#ofsharedterm[i,j]/Min(rule[i].numberofterm,rule[j].numberofterm)))Solving for the number of shared terms:

#ofsharedterm[i,j]=Min(rule[i].numberofterm,rule[j].numberofterm)*(1-D[i,j])

Then for the number of shared terms between the first rule, 1, and the rule j is,

#ofsharedterm[1,j]=Min(rule[1].numberofterm,rule[j].numberofterm)*(1-D[1,j])

Thus for the first rule, the total number of shared terms is obtained by summing over all rules j(including 1):

$$\Sigma\#ofsharedterm[1,j]=\sum_{j=1}^{n}\text{Min}(rule[1].numberofterm,rule[j].numberofterm)*(1-D[1,j])$$

For the entire rule base, the total number of all shared terms can be defined as:

Total number of shared terms =

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]$$

Expanding this sum, the total number of shared terms in a rule base can be described as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]=\sum_{i=1}^{n}\sum_{j=1}^{n}(\Sigma Min(rule[i].numberofterm,rule[j].numberofterm)*(1-D[i,j]))$$

Let the average # of terms in a rule of a rule base be denoted as AvgTerms,

then AvgTerms=total # of terms/total # of rules

Assume:

$Min(rule[i].numberofterm,rule[j].numberofterm) \cong AvgTerms$

Then I can rewrite the formula for sum of all shared terms as:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]=\sum_{i=1}^{n}\sum_{j=1}^{n}AvgTerms*(1-D[i,j])$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}AvgTerms - \sum_{i=1}^{n}\sum_{j=1}^{n}AvgTerms*D[i,j]$$

If there are N rules in a rule base, and since AvgTerms is a constant,

$$=AvgTerms*(N^2 - \sum_{i=1}^{n}\sum_{j=1}^{n}D[i,j])$$

In other words,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]/AvgTerms=N^2- \sum_{i=1}^{n}\sum_{j=1}^{n}D[i,j]$$

where

$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]$ = Total # of shared terms

Since I assumed that AvgTerms=total # of terms/total # of rules, the equation can be rewitten as :

=Total # of shared terms*(total # of rules/total # of terms)

=(Total # of shared terms *total # of rules)/total # of terms

Since the ratio total # of shared terms/total # of terms is the rule base homogeneity and total # of rules = N, I can write

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\#ofsharedterm[i,j]/AvgTerms=Rule\ Base\ Homogeneity*N$$

$$Rule\ Base\ Homogeneity*N= N^2- \sum_{i=1}^{n}\sum_{j=1}^{n}D[i,j]$$

Dividing both sides by N,

$$Rule\ Base\ Homogeneity= (N^2- \sum_{i=1}^{n}\sum_{j=1}^{n}D[i,j])/N$$

$$\text{Rule Base Homogeneity} = N - \left(\sum_{i=1}^{n}\sum_{j=1}^{n} D[i,j]\right)/N$$

Thus, my approximation is:

∴ Rule Base Homogeneity= # of rules - Sum of all distances/# of rules

For the three real rule bases considered, Table 3-1 shows the Lexical Homogeneity values.

Table 3-1 Lexical Homogeneity Values of Rule Bases

Rule Base Name	Number of Rules	Sum of All Distances	Homogeneity
"BACTEREM"	268	68318.3	13.08
"CONTRACT"	147	15444.9	41.93
"ADVICE"	172	7054.81	130.98

The BACTEREM rule base is a KBS for medical diagnosis which was developed for use in hospitals in Israel. It has 268 rules and was built using the VP-Expert shell. "BACTEREM" rule base is least homogeneous. The CONTRACT rule base is a KBS which is used for selecting contractors for construction work. It has 147 rules and was built using the VP-Expert shell. "CONTRACT" rule base is somewhat homogeneous. The ADVICE rule base is a KBS for advising foreign students in selecting American graduate schools. It has 172 rules and was built using the VP-Expert shell. "ADVICE" rule base is very homogeneous.

4. Conclusion and Further Study

I suggest that the concept of lexical homogeneity can be a good measure to predict the difficulty in maintaining a rule base. I was intuitively convinced that lexically homogeneous rule bases would be easier to modify rules since it must be easier for the maintainer to understand the structure as well as the semantics of a rule base; human cognition can understand things that are less complex. However this claim needs to be proven through a rigorous experiment using human subjects who will conduct maintenance works on different rule bases whose lexical homogeneity value is different. This human experiment is for future research.

References

[1] F. Coenen and T. Bench-Capon, *Maintenance of Knowledge-Based Systems*, Academic Press, 1993.