

Representation of data mining results in natural language

The goal of data mining is to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner [HMS 01]. The problem can arise when the data owner (e.g. physician) get set of association rules true in the analysed data – it can be simply too much of formalised information.

The association rules can be more understandable when they are presented in natural language. This way can be results of data mining better disseminate among users not familiar with data mining procedures.

The possibility of presentation of results of association rules produced by GUHA procedures is discussed in [Ra 97]. This idea were further developed and an experimental system AR2NL was implemented. The system AR2NL converts association rules into natural language.

The current version of the AR2NL system deals with association rules of the form of founded implications produced by the procedure 4ft-Miner. A simple example concerns patients from the medical project STULONG. The association rule

$$\text{Responsibility In a Work}(\text{managerial worker}) \Rightarrow_{0.93, 37} \text{Education}(\text{university})$$

related to the 4ft table

Normal group	Succedent	\neg Succedent	Total
Antecedent	37	3	40
\neg Antecedent	173	150	323
Total	210	153	363

was converted into several versions of natural language presentation, see also here.

Version No. 1:

93 % (viz. 37) of the patients show this relation: if the patient works as a senior manager and has reached university education, then he also mainly sits in his job.

Version No. 2:

Patient that works as a senior manager and has reached university education also mainly sits in his job. This fact is confirmed by 93 % (viz. 37) of the patients.

Version No. 3:

Patients that work as senior manager and have reached university education also mainly sit in their job. This hypothesis is confirmed by 37, i.e. 93 % of the patients.

Version No. 4:

37 (i.e. 93 %) of the patients that work as senior manager and have reached university education also mainly sit in their job.

Version No. 5:

It is characteristic for the patients that work as senior manager and have reached university education that they also mainly sit in their job. (This rule is confirmed by 37, i.e. 93 % of the patients.)

The core of the system AR2NL are several linguistics tools developed by P. Strossa. The system AR2NL is implemented by Z. Černý. For more details see publications [SR 02], [SR 03], [RSC 03], [St 04]. The first version of AR2NL was implemented in the frame of master thesis [Ce 03]. The current version of the system AR2NL is embedded into medical project STULONG.

The current research concerns possibilities of presentation of further types of association rules produced by the procedure 4ft-Miner [St 04]. Also possibilities of automatic production of various summaries and analytic report are investigated, see also [Ra 97].