Chapter 5

Problem solving type task: Discovering trends in the human resources for research and technology

This chapter presents an investigation into the capability of the CILT-MAL system to perform a problem solving type task. This task concerns trend analysis with regards to the quality of human resources, in scarce disciplines of specialisation for research and technology in the current science and technology system of South Africa. Where scarce disciplines of specialisation are defined as *those disciplines in which participating organisations experience difficulty to find suitable staff.* Data supporting three of the surveys, conducted as part of the NRT Audit, were used for this investigation.

The goal of the intelligent data analysis presented here is to show that a CILT-MAL system can be used successfully for problem solving type tasks. The chapter is organised as follows: Section 5.1 outlines the task on hand; Section 5.2 discusses the data pre-processing step, followed by Section 5.3 that introduces the initial exploratory results. Section 5.4 explains the experimental method and evaluation criteria used for the analysis. The actual learning by the system is presented in Section 5.5, followed by the validation and knowledge fusion episodes in Section 5.6. The chapter concludes with, Section 5.7, a discussion of the results produced by the CILT-MAL system and Section 5.8, the conclusions drawn from the intelligent data analysis.

5.1   Task description

The following three surveys, conducted during the NRT Audit, were used to investigate the quality of human resources, in scarce disciplines of specialisation, namely:

- Data from the Survey on human resources and skills in science, engineering and technology [DACST 1998], focusing on human resource delivery to the current science and technology system. A census of all the participants in the higher education system was conducted to obtain this data.

- Data from the Survey on scholarship, research and development [DACST 1998], focusing on the human resources related to the research and technology output of the current science and technology system. Data supporting this survey was obtained from a representative sample of organisations.

- Lastly, data from the Survey on the technology base of the South African business sector [AMI 1998], focusing on the economic relevance, effectiveness and efficiency of the human resources employed by the business sector. Data supporting this survey was also obtained from a representative sample of organisations

5.2   Data pre-processing

The data collected by the three surveys were carried on different entities in the NRT Audit Data Warehouse. The entities are related to one another as shown in Figure 8.
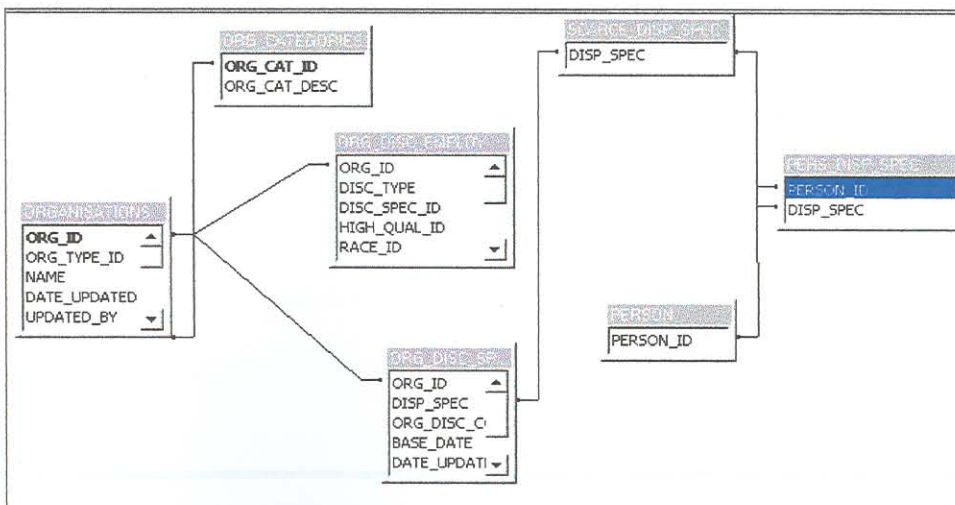

Figure 8 Human resources data structure

The data pertaining to this analysis were extracted from the different entities and integrated into one data set, as discussed next. Each organisation or functional unit was represented by one instance on the *Organisation* entity. The unique identifier for an organisation and the organisation's type describes this instance. Instances on the *Org_Disp_Employ* entity describe the number of employees within an organisation in terms of the employee's discipline of specialisation, gender, race and highest qualification. The *Org_Disp_Employ* entity contains data supporting only the survey of the technology base of the South African business sector, i.e. data covering the private sector. The *Scarce_Disp_Spec* entity listed all the disciplines in which participating organisations experience difficulty to find suitable staff for, as indicated during interviews or on returned questionnaires. Since this task investigated the quality of human resources, in scarce disciplines of specialisation, only data pertaining to these scarce disciplines were included. The *Person* entity contains data describing employees in the public sector including universities, technikons and scientific councils. These instances were described by their gender, race and highest qualification. The *Person_Disp_Spec* entity contains the specific disciplines of specialisation for all the employees on the *Person* entity. Again, only data associated with people in the scarce disciplines, as indicated by participating organisations, were included. Table 10 lists these scarce disciplines.

Table 10 Scarce disciplines of human resources in research and technology

Applied Biological Sciences	Applied Mathematical Sciences	Applied Physical Science	Engineering
Agricultural economics	Information and computer science	Chemical sciences	Engineering sciences
Agriculture	Computer networks	Analytical sciences	Chemical engineering
Food science & technology	Information systems	Chemistry, general	Civil engineering
Forest science	Programming systems	Applied chemistry	Electronical engineering
Horticulture		Polymers	Electronic engineering
Wood science		Textiles	Mechanical engineering
Biochemistry		Material sciences & technology	Manufacturing & process technology
Viticulture			Production management
Oenology			

Table 10 shows that the scarce disciplines identified by the organisations predominantly reside under the applied science categories. To ensure sufficient coverage for the disciplines of specialisation, the disciplines were grouped under four major applied science categories, namely applied biological sciences, applied mathematical sciences, applied physical sciences, and engineering.

The instances in the integrated data set were described by the following attributes:

Gender: male or female,

Race: advantaged or disadvantaged,

Discipline of specialisation: applied biological sciences (ABS), applied physical sciences (APS), applied mathematical sciences (AMS) or engineering,

Highest qualification: diploma, 4 year degree, or an advanced degree; and

Organisation type: private sector, university, technikon or science council.

Table 11 Sample data from the employee-profile data set

Highest qualification	Race	Gender	Discipline specialisation	Organisation type
Postgrad	A	M	ABS	Private
Diploma	A	M	Engineering	Private
Degree	A	M	ABS	Private
Diploma	D	M	Engineering	Private
Postgrad	A	F	APS	Private
Degree	A	M	Engineering	Technikon
Postgrad	A	M	Engineering	University

The employee-profile data set contained 1422 instances. The aim of the learning task was to identify trends in this employee-profile data set of employees in scarce disciplines.

5.3 Initial exploratory results

The learning task, to identify trends in this employee-profile data set of employees in scarce disciplines, was not a typical supervised learning task. Recall, from Section 2.5.1 that, during supervised learning the learning element is presented with a training set which describes each instance by both the predicting attributes, as well as its predicted attribute, i.e. class. The task of the learning element is to generate a class description for each domain value of the predicted attribute. In contrast, this learning task described here involved the identification of unknown trends as well as generating descriptions for them. The performance element does not only have to generate a class description for each domain value of the predicted attribute, but first had to identify the best possible predicted attribute from the set of attributes. This is an example of an unsupervised induction problem, as defined in Section 2.5.1, where both the classes and their corresponding descriptions are unknown.

The CN2 induction algorithm was adapted from a supervised method to an unsupervised method. This enabled the system to determine the strongest and most interesting set of

predicting attributes and the associated predicted attribute. CN2 was adapted in the following way: *Given k attributes, the algorithm executes k times. In each case with a different attribute playing the role of the predicted attribute. This results in k different classifiers, each designed to accurately predict the predicted attribute as a function of the remaining predicting attributes* [Langley 1998]. From the results obtained by the execution of the above-mentioned procedure, the knowledge engineer, after consulting the human expert, identified the organisation type as the strongest and most interesting predicted attribute. The strength of the classifier was measured by means of its degree of correlation i.e. overall rule set accuracy as determined by the CN2 evaluation function. In addition, the human expert evaluated the interestingness of the generated rule set. The employee-profile data set with organisation type as predicted attribute was chosen to be further processed.

5.4   Experimental method and evaluation criteria

From the employee-profile data set, randomly chosen training, test and validation sets were generated. Following Theron (1993), the training set contained 70% of the available instances and the test and validation sets comprised of 15% respectively. Only five of the 1422 instances describe the science council organisation type. Due to this low occurrence, these instances were removed from the data set.

The CILT-MAL system consisted of two machine learners, CN2 and C4.5, together with a human learner. Preliminary training showed that the BRAINNE learner was unsuitable due to its low fidelity. Recall that the fidelity refers to the extent in which the accuracy of rule extraction matches that of the original trained neural network. Here, the neural network was 82% accurate, whilst the BRAINNE system was 73% accurate.

Each of the three participating learners was presented with the full training set of 993 instances and the test set of 212 instances. The learners proceeded by executing their individual inductive learning episodes as described in Section 2.4.1, followed by the evaluation step. All trends were evaluated by the CN2 evaluation function, and converted to the performance measures for problem solving type tasks. These performance measures

were determined in collaboration with the human expert during preliminary data analysis and are defined as follows:

- **Strength**: *The percentage positive coverage of a trend is defined as the total number of positive instances covered by the rule divided by the total number of instances.* For example, if R_1 classifies 95 instances correctly and 4 instances incorrectly the strength of the trend will be 95.96%. The trend will be deemed strong when the calculated strength is higher than the average percentage positive coverage as calculated over all the trends in the rule sets generated by the learners, which are participating in the CILT-MAL system.

- **Interestingness**: *The difference between the percentage positive coverage of a trend over the different classes.* For example, the training set has 754 instances belonging to class A, 66 belonging to class B and 173 belonging to class C. R_1 covers 95 of the class A instances, two of class B's and two of class C's instances positively. The percentage positive coverage for each class is thus 13%, 3% and 1% respectively. A significant difference between the relative coverage exists. This implies that the identified trend (R_1) is a common trend of class A when compared to the other classes, which makes it interesting and worthwhile to investigate further. As a pre-requisite the highest percentage positive coverage must be that of the assigned class.

- **Consistency**: *Consistency is determined in terms of how consistent the strength of a certain trend is when it is evaluated against different data sets.* For example, the strength of trend R_1 measures 96% on the training set, 100% on the test set, and 93.3% on the validation set. The discrepancy value of the trend is Δstrength on the test and validation sets, calculated as follows:

  $100 - 93.3 = 6.7$,

  a low discrepancy value indicates a consistent trend. This measure ensures that a trend is not just a one-off occurrence within one data set, but occurs with similar strength in all the data sets.

These performance measures differ significantly from that of the classification task presented in Chapter 4. Recall that the goal of the classification task is to generate rule sets that can accurately predict the class of unseen instances. Therefore, performance is measured

in terms of overall rule set accuracy. However, the main goal of a problem solving type task is to find strong, consistent and interesting individual trends within a given data set. The significance of the trends as a disjunctive set with predictive power, measured by means of the overall rule set accuracy, is therefore irrelevant. The unique performance measures, as defined above, were determined by analysing the results of the individual inductive learning episodes, in collaboration with the human expert. Pazzani (2000) is of the opinion that most KDD papers contain unfounded assumptions about "interestingness". For example, the paper, "Mining the most interesting rules" by Bayardo and Agrawal [Bayardo *et al* 1999] presents an algorithm that searches the space of association rules with metrics involving the confidence of the rules. However, the paper does not show that any of these metrics correlates with user judgements of what is interesting [Pazzani 2000]. Therefore, when the performance measures for this specific problem solving type task were defined it was done in collaboration with the human expert. Bollacker *et al* (2000) accomplished this type of collaboration, when determining the interestingness of new scientific literature on the web, by calculating the interestingness of a new paper as a weighted sum of the relatedness between the new paper and existing papers in a user's profile. When the interestingness is greater than a certain threshold value the system, CiteSeer, recommends the new paper to the user. The user could then manually adjust the calculated relatedness of the new paper added to his/her profile [Bollacker *et al* 2000].

The remainder of this chapter describes the learning as it occurred within the CILT-MAL system. The chapter concludes with a discussion of the resulting trends.

5.5 Learning of the problem solving type task by the combined machine-human learner team

This section describes the individual and co-operative learning of the combined machine-human learner team. Two inductive machine learners, CN2 and C4.5 and a human learner participated in the combined machine-human learner team.

5.5.1 Individual learning phase

The three learners, i.e. the CN2, C4.5 and human learner, received the full set of 993 training instances and proceeded with the individual inductive learning episode. The

knowledge engineer, as introduced in Section 2.5.7.2, transferred the knowledge acquired by the human learner into a decision list consisting of 20 trends.

The evaluation steps, using the training and test data sets, followed the inductive learning episode. The trends generated by each learner with their resulting evaluation measures were added to the learner's individual knowledge bases. Table 12 summarises the results of the individual learning phase.

Table 12 Strengths and interestingness of trends after the individual learning phase

Learner	Strength		Interestingness		Total	High quality trends
	Strong	Weak	Interesting	Uninteresting		
CN2	15	22	15	22	37	9
C4.5	8	0	4	4	8	4
Human learner	14	6	10	10	20	6

The average strength of the trends, evaluated against the test set, was 49.7%. The trends had an average consistency of 86.4%. The CN2 learner produced the highest number of trends, followed by the human expert and then C4.5. The C4.5 learner produced eight trends during the individual inductive learning episode, compared to CN2's 37 trends. The CN2 learner produced nine strong and interesting trends, compared to C4.5's four strong and interesting trends and the human learner's six. The average class coverage of the trends, generated by the C4.5 learner, for the different classes were, 15% for a trend describing private organisations, 0% for a trend describing technikons and 21% for one describing universities. On the other hand, the CN2 learner's average class coverages for the respective types of organisations were 7%, 3% and 7%.

This indicated how the differences in the performance elements, i.e. a general-to-specific search directed by information theoretic entropy and significance measures of CN2, as described in Section 2.5.4, compared to a greedy search directed by gain ratio criteria and error-based pruning of C4.5, as described in Section 2.5.5, of the two learners influence the

way they perceive the problem domain. The CN2 learner generated the highest number of rules with a low average class coverage, indicating more specific trends compared to the C4.5 learner's lower number of rules with a significant higher average class coverage, indicating more general trends.

For example, the following high quality trend was generated by CN2:

```
IF      Highest_qualification = Diploma
   AND Race = D
THEN  Org_type = Private [95 2.50 0.50]

                FIRED?
ACTUAL CLASS Yes      No     Accuracy
     Private     95      659     12.6 %
Not Privat    3      236     98.7 %
Overall accuracy: 33.3 %
```

compared to a high quality trend generated by C4.5 for the same class

```
IF     Highest_qualification = Diploma
THEN  Org_type = Private  [297 24 2]

                FIRED?
ACTUAL CLASS Yes      No     Accuracy
     Private    287     467     38.1 %
Not Privat   26     213     89.1 %
Overall accuracy: 50.4 %
```

Note the difference in coverage, for the "Private" class, 95 instances were covered by the rule produced by the CN2 learner compared to 297 instances covered by the trend discovered by the C4.5 learner.

The human learner created more or less one trend for every possible attribute test combination. This limited the coverage of the trends, similar to that of the CN2 learner, creating strong uninteresting trends.

5.5.2 Co-operative learning episode

The three learners then engaged in co-operative learning, querying the knowledge bases of the other members for high quality trends related to their own low quality trends, as described in Section 2.4.2.

The problem solving data generation approach, as introduced in Section 2.4.2, was executed next. Using this approach, only one single new training set based on all the high quality trends contained in the NewRule list was generated. The coverage of the high quality rules that need to be learned in the new training set has doubled and the remaining instances, which were not covered by the new high quality rules, were left as is. Therefore, instead of generating one strongly biased data set for each high quality rule, one single data set was generated with twice the number of instances covering each high quality rule on the NewRule list. This new training set was merged with the original training set and presented to the learners.

5.5.2.1 The CN2 co-operative learning episode

The CN2 learner obtained three high quality trends from the human learner's knowledge base and four from the C4.5 learner's knowledge base. These seven high quality trends were added to the CN2 NewRule list and the 28 low quality trends were removed from CN2's knowledge base. Table 13 contains six of the 28 low quality trends generated by the CN2 learner:

Table 13 Low quality trends generated by the CN2 learner

Rule	Highest qualification	Gender	Race	Discipline specialisation	Organisation type	Positive coverage
16	Diploma	Male		AMS	Technikon	0%
19	Diploma	Female	Adv	ABS	Technikon	0%
23	Diploma	Male	Adv	Engineering	Technikon	25%
24	Diploma	Male	Adv	Engineering	Technikon	15.8%
26	Diploma	Male	Adv	APS	Technikon	0%
29	Diploma	Female		AMS	University	0%

These trends were in conflict with the C4.5 high quality trend, namely

```
If    Highest qualification = Diploma

Then  Organisation type = Private    [287 24 2],

strength 92.6% (on test set).
```

The C4.5 high quality trend indicates that, if a person's highest level of qualification is a diploma, then the person will most likely be working in the private sector. The six low quality trends in Table 13, generated by CN2, indicated that, if a person's highest level of qualification was a diploma, then the person would most likely be working at a university or technikon. These trends are in direct conflict.

The data generator proceeded to generate 313 new instances to cover the C4.5 high quality trend on the NewRule list. Of the 313 instances, 287 were of the private organisation type, 24 were technikons and 2 universities, similar to the original distribution. The attribute values were distributed in such a manner that the mean values and variances were the same as that of the original 313 instances. The 313 new instances were added to the original training set. Similarly, data was generated for all the other high quality trends on the NewRule list and added to the training set. The CN2 learner was then presented with this

new training set consisting of 1 758 instances, including the 993 instances of the original training set. CN2 then generated the high quality trends shown in Table 14.

Table 14 Trends generated by the CN2 learner during the co-operative learning episode

Rule	Highest qualification	Gender	Race	Discipline specialisation	Organisation type	Positive coverage
1	Diploma		Disadv	APS	Private	100%
2	Diploma	Male		ABS	Private	100%
4	Diploma		Disadv	Engineering	Private	100%
5	Diploma	Female			Private	92%
6	Diploma	Male	Adv		Private	90%

CN2 generated five individual high quality trends. The five rules can be combined into a single rule, R_{new} as follows:

$$(R_1 \cup R_2 \cup R_4 \cup R_5 \cup R_6) \implies R_{new}$$

where R_{new}: If   Highest qualification = Diploma

and   Discipline specialisation in (APS,ABS,Engineering)

Then   Organistion type = Private.

R_{new} resembles the C4.5 high quality trend on the NewRule list:

If   Highest qualification = Diploma

Then   Organisation type = Private.

R_{new} differs from the trend on the NewRule list, in that the CN2 learner excluded the applied mathematical science's (AMS) discipline from the concept description. On further investigation, it was found that only seven of the 313 instances covered by the trend were from AMS. This low coverage was the reason why CN2's search algorithm excluded this discipline.

After completing the co-operative learning episode, the CN2 knowledge base consisted of nineteen high quality trends. The CILT learning system subsequently applied a rule-pruning algorithm to simplify the knowledge base. As discussed in Section 2.2.2, when simplifying the knowledge bases of a problem solving type task, duplicate trends are removed. Any further simplifications, for example the combination of trends, depended on the user's needs and not quantitative measures. In this study, all the duplicate trends were removed from the CN2 knowledge base.

5.5.2.2 The C4.5 co-operative learning episode

During the individual learning phase, C4.5 produced eight trends of which four were strong and interesting. The uninteresting trend,

```
If   Highest qualification = Degree
Then Organisation type = Private,
```

was not removed from the C4.5 knowledge base. This was due to the fact that this trend had the highest percentage class coverage, 69%, for technikons, with a percentage coverage discrepancy of 26% and 32% for private sector and universities respectively. This indicates a strong trend for technikons, although the classification was incorrect. Implying that the majority of academic staff, employed by technikons in South Africa, highest level of education is a four-year degree, although most of the four-year degree employees are to be found in the private sector. The remaining three low quality trends were removed from the C4.5 knowledge base since no other high quality trends related to these low quality trends could be found.

5.5.2.3 The human learner co-operative learning episode

The human learner generated six high quality and 14 of low quality trends during the individual learning phase. During the co-operative learning episode six high quality trends

were added to the NewRule list. These trends related to the fourteen low quality trends and were selected from both the CN2 and C4.5 knowledge bases.

The human learner, with the assistance of the knowledge engineer, removed seven trends, including three of low quality, from the knowledge base and added one from the NewRule list that corresponded to the combined criterion of the seven individual trends. Four misconceptions were removed from the knowledge base and replaced by two trends from the NewRule list. A high quality trend contained in the NewRule list that overlapped with two low quality trends, was added to the knowledge base. The two low quality trends were removed. This resulted in a human learner knowledge base consisting of seven high quality trends. Table 15 summarises the results of the co-operative learning episodes.

Table 15 Results of the co-operative learning episodes

Learner	Strong	Interesting	Total	Number of rules in knowledge base
CN2	18	13	13	13
C4.5	5	5	5	5
Human expert	7	6	7	6

Notice that at this time the consistency has not been determined since it can only be calculated after the validation episode. This is due to the fact that strength measurements against both the test and validation set are necessary for the calculation.

5.6 Validation episode and knowledge fusion

During the validation episode, the final knowledge bases of the team members participating in co-operative learning were evaluated against the validation data set. The third performance measure, namely, consistency was used at this time. This was done to ensure that the high quality trends were not co-incidental, but consistent over the two randomly chosen unseen data sets namely the test and validation set. The consistency of the trends were calculated at this point and not earlier, because only at this point were strength measurements available as evaluated against the test and validation set that could be

compared. The average consistency of all the trends in the high quality knowledge bases was 92.8%. Trends, with a consistency higher than the average, are deemed to be consistent, as defined in Section 5.4. However, the trends with a lower than average consistency need to be further investigated. Due to the fact that the test and validation sets are smaller than the original training set, a one-instance difference in coverage, of a trend with low class coverage, can make a significant difference in the consistency of the trend. For a well-informed decision, both the actual positive and negative coverage as well as the percentages should be taken into consideration. For example the C4.5 trend:

```
If      Highest qualification = Postgrad
   and     Discipline specialisation = APS
   Then  Organisation type = Private,
```

with a (private, technikon, university) [59 1 11], [13 0 5] and [9 0 1] coverage over the training, test and validation sets respectively, had a 17.8 discrepancy value, calculated as follows:

$$|13/18*100 – 9/10*100| = 17.8$$

The consistency value of the trend is therefore, 100-17.8 = 92.2%. However, the following trend was found to be more consistent than the above-mentioned trend although its consistency value, calculated as follows, was lower:

```
If    Highest qualification = Postgrad
   and    Race = Disadvantaged
   and    Gender = Male
   and    Organisation type = Private,
```

with a [9 0 4], [1 0 0] and [3 0 1] coverage over the training, test and validation sets. The discrepancy value 25 was calculated as follows:

$$|1/1*100 – 3/4*100| = 25.$$

The consistency value of the trend is therefore, 100-25 = 75%. The coverage of this trend is lower than for the previous case and therefore the only one instance that was misclassified resulted in a lower calculated consistency.

Finally, the knowledge bases of the individual team members were fused together into one integrated knowledge base, consisting of 21 trends, which contained the results of the team effort. The team knowledge base was then pruned for duplicates, removing six trends, resulting in a 15 high quality trend team knowledge base.

5.7   Discussion

This section highlights the main findings as contained in the team knowledge base. The human expert's perspectives were obtained by an interview as well as by consulting the synthesis report. The trends, as contained in the knowledge bases, in some instances supported the findings of the human expert while contradicting the expert in other cases. There were newly discovered trends of which the human expert was not aware.

The knowledge acquired by the learning system supported the human expert on the following trends. Firstly, that there is a significant difference in the type of organisation where a person will be employed depending on his/her highest qualification. Persons with a four-year degree or diploma, as defined in Section 5.2, are predominantly employed by the private sector. Within the private sector and technikons the percentage of employees with diplomas are identical. However, the technikons dominate with a significant higher percentage of employees with a four-year degree compared to the universities and private sector that are similar. This confirmed the human expert's opinion that, within a technikon, a person with a four-year degree can obtain a more senior position.

Secondly, that the lack of presence of postgraduate engineers in the private sector is due to the weakness of the South African industry's own process development initiatives. The knowledge base indicated that universities, as tested against the test set, employ 90% of the advantaged, male, engineers. The human experts were of the opinion that the South African industry prefers to buy in processes by licensing (AMI 1998). However, postgraduate engineering courses train people for product development, without the South African industry having the required maturity to compete internationally with new products. This indicates a mismatch in the needs of the industry and the focus of the higher education

system. However, disadvantaged engineers with a Masters or Doctorate are absorbed by the private sector. The knowledge base indicates that 75%, as tested against the test set are employed within the private sector. This could be due to the imbalance in the production of engineers from disadvantaged backgrounds. This imbalance is most prominent in this field. This is amplified by the fact that the higher education system cannot compete with the salaries offered by the private sector and therefore the trend becomes more prominent.

Lastly, that the working conditions in the information technology industry are the least suited to that of a working mother. Therefore, females with advanced degrees in applied mathematical sciences, for example information technology, prefer working at universities, even though there are plenty of opportunities for employment within the private sector. This was confirmed by a similar trend in the knowledge base.

Some of the opinions of the human expert were, contradicted by trends contained in the knowledge base, highlighting preconceived biases of the human expert. For example, the strongest trends based on gender were for females in applied physical science, females with degrees in the applied biological sciences and females with Masters or Doctorates in the applied mathematical science, as listed below.

```
IF    Gender = F
   AND Discipline_specialisation = APS
THEN  Org_type = Private  [54 2 1]


IF    Highest_qualification = Degree
   AND Gender = F
   AND Discipline_specialisation = ABS
THEN  Org_type = Private  [31 0 0]

IF    Highest_qualification = Postgrad
   AND Gender = F
   AND Discipline_specialisation = AMS
THEN  Org_type = University  [0 0 2]
```

These trends confirmed the human expert's opinion that females in the applied physical science work in the private sector. These females are responsible for most of the routine work, such as conducting tests in analytical laboratories and quality control. However, the human expert was convinced that females in the applied biological sciences would not be employed by the private sector since biotechnology within South Africa is dominated by the government sector. However, the trends in the team knowledge base contradicted the expert and indicated that the food processing industry is a major employer of females trained in this discipline.

Some trends discovered by the CILT-MAL system were new to the human expert as they were previously unknown and potentially useful information was discovered by the intelligent data analysis tool. For example, the trend that employees with an advanced degree, i.e. a Masters or Doctorate, in the applied physical sciences were strongly absorbed by the private sector compared to, for example engineers. This is due to the fact that the scarce applied physical sciences discipline was dominated by chemistry. The South African chemical industry is recognised as an international competitive industry, competing with its own products. Because of its sophistication, this industry employs human resources similarly distributed to corresponding organisation types in Germany and the Unites States (Synthesis Report). This implies that the chemical industry can offer postgraduate qualified candidates better career opportunities and therefore becomes their major employer.

5.8   Conclusion

This chapter presented a problem solving type task, executed by a CILT-MAL system, applied to a real-world scenario. Section 5.1 described the task that concerned the discovery of trends in the human resources for research and technology, in the scarce disciplines of specialisation, by means of co-operative learning in a multi-agent learning environment. Section 5.2 discussed the integration of the data from the different surveys into a single data set. Section 5.3 explained how the CN2 learner was adapted from a supervised to an unsupervised learning algorithm. The experimental method and evaluation criteria were introduced in Section 5.4. Section 5.5 introduced the machine-human learner combination team and showed how co-operative inductive learning manifested within the MAL system.

Next, the knowledge fusion episode was discussed in Section 5.6. The chapter concludes with a discussion of the major trends discovered during the co-operative learning episode.

Results indicated that a CILT-MAL system could be successfully applied to a real-world problem solving type task. However, for the CILT-MAL system to be successful the following had to be defined; a new set of performance measures; an alternative method for the data generator to generate data with; as well as a trend pruning algorithm. Once again, the results as contained in the team knowledge base confirmed some findings of the experts, but also showed where the findings were wrongfully made, making it a valuable learning experience to both the experts and the learners.

Problem solving type tasks are more complex than classification tasks, due to the multiple steps involved. Therefore, executing this type of task becomes a more time consuming process. The CILT-MAL system could be a valuable tool assisting the human expert during his/her execution of the task, saving the expert time as well as verify his/her findings.