DATA QUALITY MEASUREMENTS FOR KDD

José Hernández-Orallo
Universitat Politècnica de València
Dep. de Sistemes Informàtics i Computació,
C/ de Vera s/n E-46022, València, Spain
E-mail: jorallo@dsic.upv.es

Francisco Alamañac Felipo
Lignotock, S.A.
Departamento Técnico
A3-Km 332, E-46930 Quart de Poblet, Vcia., Spain
E-mail: francisco.alamanac@sommer-allibert.com

Résumé. Ce papier introduit *mesures* de qualité de données conçues spécialement pour la découverte de connaissance dans bases de données (KDD). D'autres méthodes, comme les auditoriats, le nettoyage et les *métriques* de qualité de données, bien que utiles, sont des procès a posteriori qui sont incapables de sélectionner le meilleur échantillonnage pour KDD. Une mesure détaillée permet de choisir une partie des données originelles avec un accroissement de la qualité et un décrement du volume de l'entrée significatifs des méthodes de KDD, améliorant, donc, leur précision et efficience. En particulier, nous introduisons un cadre de mesure qui peut être ajouté à virtuellement quelque système d'information sans une surcharge appréciable pour l'utilisateur. Quelques résultats préliminaires suggèrent que la mesure de qualité représente un champ inexploré où la technologie KDD peut encore améliorer dans un ordre de magnitude avec les techniques actuelles d'exploitation de données.

Mots Clefs: Qualité de Données, Exploitation de Données, Découverte de Connaissance dans Bases de Données (KDD), Systèmes d'Information, Atrophie de Composants.

Abstract. This paper introduces data quality *measurements* especially designed for data mining. Other methods, like data audits, data cleansing and data quality *metrics*, although useful, are all off-line processes which are unable to *select* the best sample or data subset for data mining. A fine-grained quality measurement allows a selection of the original data with a significant quality increase and size decrease of the entry of data mining methods, so improving their accuracy and efficiency. In particular, we introduce a measurement framework that can be easily added to virtually any information system without appreciable overload to the user. Some preliminary results suggest that quality measurement techniques represent an unexploited field where KDD technology can still improve in an order of magnitude with current data mining techniques.

Keywords: Data Quality, Data Mining, Knowledge Discovery in Databases, Information Systems, Use Frequency, Component Atrophy.

1. INTRODUCTION

Recently, there has been an increasing interest on the study of data quality (Redman 1996) in information systems, especially in modern data warehouses. The accuracy of vital data has been considered a major problem in many dynamical systems, from legal to financial databases, even if these systems are designed with correctness as a primary issue and, initially, they perform correctly. The main reason for this loss of accuracy in dynamical systems is time. As time goes by, more information is getting obsolete. However, this usually happens to those pieces of data which have not recently been used. This phenomenon is generally known as *atrophy*, something that happens in any physical system (mechanical or biological) but it

also may happen in logical dynamic systems, like software or information systems. Theoretically, a logical system cannot deteriorate with time, but atrophy originates because unused parts become obsolete.

Traditional methods that try to alleviate this problem, like avoidance of redundancies and "data sharing", are nowadays incorporated in most information systems. However, the quality of data still decreases as the system gets old. Data audit (Parsaye and Chignell 1993) (Wang et al. 1995b) can be performed regularly in order to re-connect the database with reality. However, they are expensive and not always successful. Another original solution to this problem is to promote a stringent use of the vital data (Orr 1998). By a stringent use we refer to a 'conscious' and contrasted use, where the piece of data is connected and compared with reality.

Despite these options, most of the information systems still lack of data quality programs, either data audits or data quality metrics. However, continuous *measurement*, even if they are not used to improve data quality is *per se* a very useful tool, because it allows the user to be aware of the reliability of the data from where many decisions and processes are based. As we will see, with the current database technology, data quality measurements are cheap, so their absence is only justified by the still scarce realisation of the importance of data quality in information systems.

Notwithstanding, in this paper we highlight another significant but neglected reason for measuring the data quality of a system. To know the data quality characteristics of an information system is essential for extracting useful knowledge from it, manually or automatically, which will be increasingly more important in the future. The advantage of this combination is that data quality measurements are easy to implement and it is not necessary (although convenient) to perform data audits.

This motivates the subject of this paper, to define models and techniques for data quality measurement appropriate for knowledge discovery from databases.

1.1. Knowledge Discovery in Databases (KDD)

KDD is defined as: "*the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" (Fayyad et al. 1996b). Data Mining is just a part of this complex process; KDD includes Data Preparation, Data Mining itself, Interpretation/Evaluation and sophisticated Visualisation tools. The whole process transforms data into knowledge, because the input of the process is extensional information from databases and the output is intensional information in a comprehensible language.

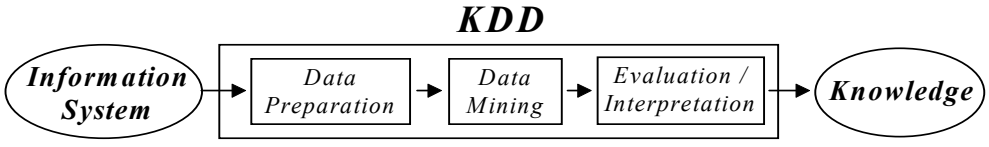

Figure 1. KDD process

Data preparation, especially its cleansing, is the most critical issue in KDD. It must eliminate redundant, inconsistent, incorrect and stale data. A single extreme erroneous or inconsistent piece of data can make the subsequent process useless. Some statistical techniques can be employed to detect these anomalous elements, but, frequently, it is impossible to discern erroneous data from surprising or crucial data, which "can actually be the key data points worth focusing on" (Brachman & Anand 1996).

2. RELEVANCE OF DATA QUALITY MEASUREMENT FOR KDD

Improvement in data quality is extremely hard and expensive to obtain. As a result, if you cannot make something good, at least it is necessary to know how bad it is.

It is generally assumed that the input of the KDD is the information system where knowledge is to be discovered. Many times it is forgotten that this knowledge will frequently be used in the same broader context where the database was generated. In other words, the *source* of the information for a database system and the *target* of the output from the KDD process is simply the same, the system environment or reality.

The relation between the environment and the database is bidirectional, i.e., there is a lot of acquisition from the reality to the database system and there are many outputs (or answered queries) from the database which are returned to reality. This process can detect acquisition (or operation) errors which logically provoke an important feedback from reality to the database, as it is shown in the left-hand side of figure 2.

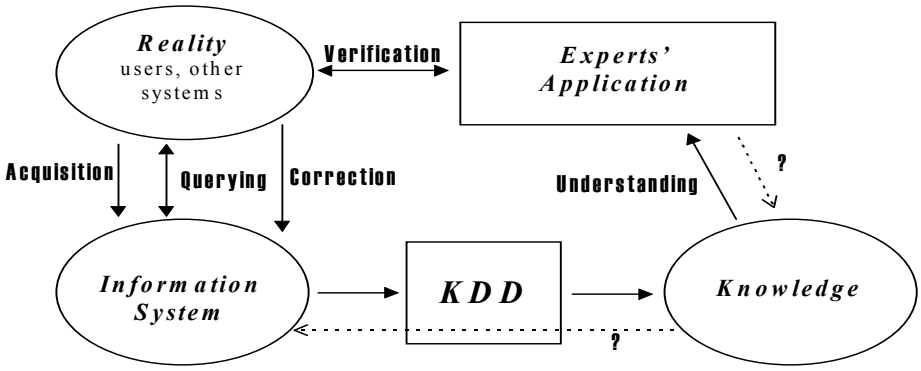

Figure 2. KDD and validation with reality

On the other side, the verification or validation of the knowledge obtained by KDD is finally contrasted with the reality and not with the information system itself. If the information system does not reflect accurately the reality, this verification would surely fail, independently of how good the KDD system could be. In (Cortes et al. 1995), it is formally shown how "*random errors and insufficiencies in databases limit the performance of any classifier trained from and applied to the database*".

Since it is impossible in general to validate completely an information system with reality, an effective and accurate measurement of the data quality of the database could be extremely profitable for KDD. However, this measurement is more useful all along the KDD process, before, during and after the data mining process.

In the data preparation stage, it is evident that the selection or sampling procedures can be highly benefited by the use of a grained or segmented information about the data quality. In this way, an information system with very low quality can still have a good KDD process if there is enough information to select the portion of data that is accurate. Preprocessing or transformation procedures can also exploit quality measures.

At first sight it seems less evident that detailed data quality information can be useful for the core of data mining itself. The data which has been selected in the preparation stage has still some quality variance and other useful information, like e.g. modification frequency and access frequency. It is reasonable to maintain this information to be processed by data mining methods. For instance, some data mining systems are based on functional dependencies; an exception can be ignored if its quality is significantly lower than the mean quality of the data which has suggested the functional dependency. Moreover, since the inductive process is extremely costly, the best 'qualified' data should be explored first.

Finally, if the data mining process cannot be modified, quality measures can be used a posteriori in the evaluation stages. For instance, reinforcement-based evaluation criteria (Hernandez-Orallo 1999) consider the *accuracy* of the examples that a model explains.

After the KDD process, and using an epistemological analogy, the left-hand side of figure 2 can be identified with perception whereas the right-hand side can be identified with cognition. Cognition is frequently useful to detect perception errors, something that most KDD systems do not reflect, a feedback that is thus represented by quotation marks. This feedback should be studied in KDD systems as an alternative to data audit. If a mismatch between knowledge and reality is detected and the KDD system determines that the knowledge failure is not caused by an error of the KDD system then the database should be revised. A similar approach is the use of ML techniques for data validation (Parsaye & Chignell 1993).

Some quality metrics are based on data samples auditing, by comparing a partial view of the *true* reality with a partial view of the database extension. Apart from the cost of this process, its efficacy highly depends on the accuracy of the expert or user which interprets the reality.

In our opinion, instead of a metric performed a posteriori, it is more reasonable to perform a continuous *measurement* about the user's interaction and satisfaction with respect to the information which is stored in the database. The accuracy of the system can be obtained by its correction frequencies rather then from the accuracy of experts.

3. FEASIBLE DATA QUALITY MEASUREMENT

Once justified the relevance of data quality *measurement*, we will evaluate which approaches introduced to date are useful (in terms of granularity), if any, and which kind of measurements can be performed without a significant overload (in space and efficiency) to the information system.

3.1. Granularity and Historical MetaData

Global metrics are not much useful for KDD. Mean database accuracy or mean database currentness are only useful to give a final value of the result of the KDD but they cannot be used to *improve* the KDD process.

More fine-grained measures, like segmented measures (Rakov 1998), where the quality values apply to homogeneous blocks or segments of data are much more useful. However, most of these approaches are oriented to obtain the quality of derived entities like views or queries. Moreover, they work under the "*assumption that (...) stored information (...) is relative static, and hence the quality of data does not change frequently*" (Rakov 1998), so they are not valid for real dynamic database systems.

Detailed measures known as "attribute-based approaches" (Wang et al. 1993) (Wang et al. 1995a) attach a 'tag' to different levels and granularities of data, to address timeliness and accuracy aspects of data. Few years ago it was argued that the additional overload did not justify the introduction of these tags. However, nowadays, almost any database system has quality tags at the level of table or even at the level of record.

Since the space for logging the tags should be minimised, a single value could be constantly updated from the previous tag and the current access characteristics and time. However, this actualisation implies a cost for each access, so it would be nice to find a compromise between long historical metadata and much too frequent access.

3.2. Measurement Framework

According to the previous considerations, we then propose an intermediate and mixed level of measurement. We will measure *global* column values and *detailed* row measurement (at the level of record).

The usual idea of adding tags to each record is not satisfactory from the point of view of independence because the database schema loses its backward compatibility since the data is not separated from the metadata. For this reason, a much modular solution consists of a parallel database which contains a historical table (*h-table*) for any other table in the actual database. Each *h-table* contains as primary key the primary key of the original table jointly with the *time* of each use. We assume that system time is precise enough to differentiate concurrent accesses. Other fields can be added to each record: kind (access, modification, creation), source (internal, external, user), stringency, a field pattern, etc. In this way, every access to any activated record in the database generates an input into an *h-table*.

For distributed information systems, there can be many de-centralised *h-tables* to record the accesses in each distributed location or access subsystem. The compromise to reduce overload is that idle times are used to process each *h-table* for computing derived metrics. These metrics are stored in a *u-table* which contains as primary key the same primary key as the original table and all the rest of derived metrics.

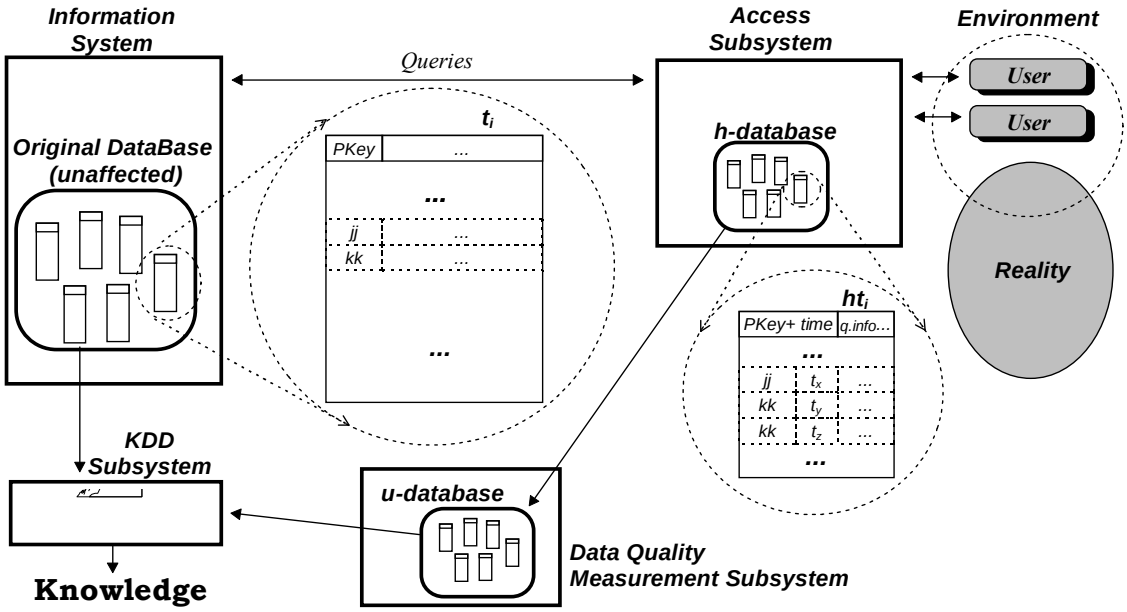

Figure 3. Hybrid Measurement Framework

Apart from this compromise between historical information and derived information, the framework becomes feasible because only stringent access should be recorded. Obviously, this obliges to determine the stringency of each interaction with the user or other systems. Non-stringent access happens when the data is not used 'consciously', or it is not visualised, or the user has not permission/obligation/intention to correct the data, or it does not know that the data is not correct, etc. In all these cases (the most), the use is not worthy and it should not be recorded. After that, it is necessary to establish a *stringency threshold* which should be tuned according to an acceptable decrease of performance of the system. However, this threshold is usually high, because "*in general, the quality of data that is not stringently used will be better than data that isn't used at all, but not much better*" (Orr 1998).

4. MODELS OF DATA QUALITY FOR KDD

Once a feasible framework for measurement is established, it is necessary to determine exactly what is to be measured, how and which derived entities can be obtained, like accuracy, 'currentness', variability, etc.

If the integrity of an information system is maintained by the DBMS, the factors which affect the quality of the system are mainly semantical. Apart from general errors and stale data, there can be still semantic redundancies and inconsistencies, due to a deficient design or an incorrect use of the schema. Among the errors we find any kind of noise: introduced by the initial source of information, skews occurred by the formalisation process, typos, internal errors and other causes. Since database systems are ever changing, time is another cause of incorrect data, and there will be more and more stable data if it is not used frequently. The typical examples of stableness are addresses, telephones, incomes, etc.

The observable phenomena to detect these problems are usage and modifications.

4.1. Models based on Usage

As we said in the introduction, there is a major factor that determines data quality, use rate. We define $U(x, t) = 1$ iff the element x has been used in time t and 0 otherwise. For convenience, we include creations and modifications as uses. We can easily define the number of times x has been used in an interval (ts,tf) as $uses\_of(x, ts, tf) = \sum_{ts \le i < tf} U(x, t)$, being $tf \ge ts$. From here we can give the following measure of use rate:

DEFINITION 4.1. *Use Rate*

Given a part x of a system S, we define its use rate as: $Use(x) = 2^{-uses\_of(x,\ tcx,\ tn)}$ being tn the current time and tcx the time of creation of x.

From here, and by using Orr's lemma of "*increase data utilisation to increase data quality*", we can introduce the simplest model of accuracy of a part x.

Model 1: $\text{Accuracy}(x) = 1 - Use(x) = 1 - 2^{-uses\_of(x)}$

This simple model, however, it is not fair with respect to the distribution of the usage. An element could be frequently used right after its creation but then it can have no use at all. In this case, the accuracy of the element is clearly lower than if the uses are more uniformly or more recently distributed. Accordingly, we use the following measure of accuracy.

Model 2: $\text{Accuracy}(x) = 2 \cdot \sum_{txc \le\ i < tn} U(x, ti) \cdot (1 - (tn - ti)/(tn - txc))$ being tn the current time, and txc the time when the component or element x was created.[1]

For instance, a component that was introduced 10 months ago, and it has had 3 stringent uses, one 8 months ago, another 7 months ago and finally 4 months ago has $\text{Accuracy}(x) = 2 \cdot (3 - 8/10 - 7/10 - 4/10) = 2.2$. It is easy to prove that, with model 2, Accuracy(x) = number of uses iff the mean of the tx_i is exactly $(t_n - tx_c) / 2$.

4.2. A Model Based on Usage and Modification

The previous model was justified for dynamic systems. However, the same reason can be used to reject it. Changes in a dynamic system usually entail modifications on its data, and each modification 'washes' the reckoning of the past use history.

We need to take into account two kinds of stringent accesses: confirmative and modifications. Confirmative accesses are equal to the uses we have defined before. Modifications are defined in a similar way: $M(x, t) = 1$ iff the element x has been modified in time t and 0 otherwise. The creation of an element will also be considered a modification. The frequency of modification access can be obtained in the following way:

[1] If stringency is a value between 0 and 1 it can be easily included in model 2 as a factor to each use.

DEFINITION 4.2. *Frequency of Modification*

The frequency of modification of a part x in an interval (t_s, t_f) is given by:

$$f_{mod}(x, t_s, t_f) = \sum_{ts \le i < tf} M(x, i) / (t_f - t_s)$$

We will assume, for convenience of implementation, the following distribution $1 - 2^{-k}$,

DEFINITION 4.3. *Probability of Future Modification*

Given a part x of a system S, the probability of modification between the time of the last modification of x (tlm) and any future instant t where $t > tlm$ can be estimated as:

$$P_{mod}(x, t) = 1 - 2^{-(t - tlm) \cdot fmod(x, tc, tlm)}$$

being tc the time of creation of x.

This probability will always strictly greater than 0 because $fmod(x, ts, tlm)$ is strictly greater than 0 (the creation is considered a modification), and $t > tlm$. Obviously, for this distribution, when $t - tlm = 1 / f_{mod}$ then $P_{Mod}(x,t) = 0.5$.

Under the lemma that "use improves quality" this probability of modification (or change) must be adjusted by the distribution of the frequency of use in the following way. For instance, if a given element used to have a low frequency of use but, since the last modification, it has undergone an increment of frequency of use, then the probability of modification decreases, since it would have been modified if it had been necessary. On the other hand, if the frequency of use has decreased, the probability of modification should be greater than the one given by $P_{mod}(x, t)$.

Consequently, we first compute the frequency of use until the last modification:

DEFINITION 4.4. *Frequency of use Until the Last Modification*

$$fulm_{use}(x) = uses\_of(x, tc, tlm) / (tlm - tc)$$

being tc the time of creation of x and tlm the time of the last modification of x. Note that this frequency can be infinite.

The accuracy is computed between the time of last modification and current time[2]:

DEFINITION 4.5. *Accuracy Since the Last Modification*

$$Aslm(x) = 2 \cdot \sum_{tlm \le i < tn} U(x, ti) \cdot (1 - (t_n - ti)/(t_n - tlm))$$

DEFINITION 4.6. *Pondered Frequency Since the Last Modification* ($t_n > tlm$)

$$Pfslm(x) = Aslm(x) / (t_n - tlm)$$

From here, we can compute the variation of frequency of use of the last modification between the period before the last modification and the period after the last modification in the following way:

DEFINITION 4.7. *Variation of Use along the last modification*

$$VU(x, tc, tlm, tn) = ((tn - tlm) \cdot pfslm(x) + (tlm - ts) \cdot fulm(x)) / ((tn - ts) \cdot fulm(x)) \text{ if } tn = ts$$

$$VU(x, tc, tlm, tn) = 1 \text{ otherwise}$$

This definition is well defined since $fulm(x)$ can never be 0 because the creation of x is considered a use. If $VU > 1$ then there are more use recently than before the last modification and if $VU < 1$ then there was more use before the last modifcation then recently. Finally, we can introduce a revised model that ponders modifications and confirmations in the following way:

***Model 3*:** $P'_{mod}(x, t) = 1 - 2^{-(t - tn) \cdot fmod(x, ts, tlm) / VU(x, tc, tlm, tn)}$

This definition is also well defined since VM can never be less than 0.

EXAMPLE 4.8. An *h-table* gives the following data about a given element x at time 100, where *C, u,* and *M* represent creation, use and modification respectively:

Time	0	4	7	8	9	10	12	17	32	34	45	56	63	65	70	75	77	78	83	84	94	96	98	99
Access	C	u	u	u	u	u	M	u	u	M	u	u	u	u	u	M	u	u	u	u	u	u	u	u

[2] Definition 4.4 could also be adjusted by a 'currentness' window that would give more value to the last uses then the first uses. For simplicity, we will only take into account the distribution of uses for definition 4.5.

If we want to find the probability of modification at instants 105, 150, and 500, first we must compute the frequency of modification until the last modification (instant 75) as $f_{mod}(x, t_s, t_{lm}) = 3 / (75 - 0) = 0.04$. The frequency of use until the last modification is $fulm_{mod}(x) = 15 / (75 - 0) = 0.2$. Then the accuracy since the last modification $Aslm(x) = 2 \cdot (9 - (116) / 25) = 8.72$, $pfslm(x) = 0.35$ and $VU = 1.186$. From here, $P'_{mod}(x, 105) = 1 - 2^{-(105-100) \cdot 0.04 / 1.186} = 0.11$, $P'_{mod}(x, 150) = 1 - 2^{-(150-100) \cdot 0.04 / 1.186} = 0.69$, and $P'_{mod}(x, 500) = 1 - 2^{-(500-100) \cdot 0.04 / 1.186} = 0.9999$.

Finally, data quality can be identified precisely with the invariant part of P'_{mod}:

DEFINITION 4.9. Quality of data under model 3

$$Q(x) = VU(x, ts, tlm, tn) / fmod(x, ts, tlm)$$

being the units of quality *time unit per modification.*

Different parameters can be added to adapt the definitions to particular information systems. It is also remarkable that all the definitions of data quality can be computed incrementally. This has made it possible for an efficient construction of *u-tables* from *h-tables* and the deletion of the latter.

4.3. Relevance

Another factor in any measurement of quality is *relevance*. Not all the data is equally important, so the quality of a system depends on the distribution between accuracy and relevance in different elements. Relevance is extremely difficult to measure. Fortunately, the use rate is related to relevance because important data is generally more frequently used, so it is implicitly taken into account in model 3. However, it is very difficult to establish a clear relation between relevance and *stringent* use.

5. EXAMPLE OF KDD APPLICATION

It is time to apply the final model for a KDD application. An information systam at the technical department of Lignotock Valencia maintains the data about 145 of its suppliers. The quality of this data is extremely important because an error in the data about the contact person, product, e-mail addresses, telephones, etc., may suppose a delay of hours or even days in an important delivery.

For this reason a quality measurement program is being added to the application. The *h-tables* and *u-tables* are constructed as they were defined in section 3. The information recorded in the *h-tables* consists of the source of the access (terminal and user), stringency, kind of access (use, creation, modification, deletion) with an optional field pattern to allow attribute granularity. The *u-tables* only include the quality, the representativeness of the quality measurement (*metaquality*) and the percentage of accesses and modifications of each user and terminal.

Theoretically, the stringency of the data accesses could be calculated from the interfaces of the forms that provide access to the information, several listings and other printed documents. In practice, for forms, the measurement subsystem only records that information which is explicitly consulted by the user and that information which gets the ‘caption’. On the other hand, only few printed documents are considered for the measurements because most of them are archival or go out from the department without feedback.

The first surprising result was that, if we activated the field pattern, the efficiency of the system was a 97% of the original one, measured in mean response time, while without the field pattern, it raised only a 98% of the original one. The second unexpected result was that data quality values were quite heterogeneous and only a 20% of the suppliers had most of their attributes with quality greater than six months per modification.

There is a log file of ISDN communications, which has registered 9443 connections since 02/09/97 to 27/01/99. It has been unsuccessfully investigated by data mining methods in order to find some relation to reduce the connection errors (these include retries). The commercial KDD package *Knowledge SEEKER IV* had been applied for the whole of data, unsuccessfully. We applied it again only to the 5% percentile of the communications of the best qualified supplier registers. The results are in figure 4:

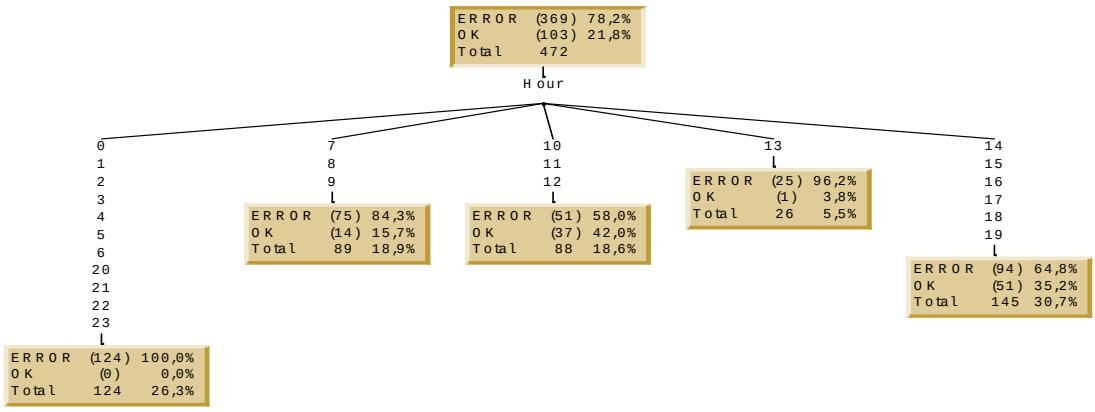

Figure 4. Split automatically generated by Knowledge SEEKER IV

The program was able to split automatically into 5 branches that show that two time zones were clearly better than the others. Some communications are being re-scheduled accordingly.

6. CONCLUSIONS AND FUTURE WORK

We have presented a measurement technique of data quality especially designed for KDD. The model is based on the use ratio and the distribution of modifications. It has been easily implemented on a suppliers database system without perceptible overload for the user. We have used it in the selection stage of a very simple data mining method. In the future we plan to exploit this *quality* metadata in the rest of phases of more complex data mining algorithms and more complex systems.

ACKNOWLEDGEMENTS

We wish to thank Lignotock S.A. from Groupe Sommer-Allibert for its permission to use its data and name, and many other facilities during our study. The data mining package Knowledge Seeker IV is copyright of Angoss International Limited, Groupe Bull.

REFERENCES

Brodie, M.L., Data quality in information systems, *Information and Management*, (3), 1980.

Brachman, R.J., Anand, T., The Process of Knowledge Discovery in Databases: A Human-Centered Approach, in (Fayyad et al. 1996a), 1996.

Cercone, N.; and Tsuchiya, M. Special Issue on Learning and Discovery in Databases, *IEEE Transactions on Knowledge and Data Engineering,* 5(6), Dec., 1993.

Cortes, C.; Jackel, L.D.; Chiang, W.P., Limits on Learning Machine Accuracy Imposed by Data Quality, in Fayyad, U.M. and Uthurusamy, R (eds.) Proceedings of the First International Conference on Knowledge Discovery and Data Mining, KDD'95, Menlo Park, Calif. The AAAI Press, 1995.

Engels, R. & Theusinger, C., Using a Data Metric for Preprocessing Advice for Data Mining Applications, in Henri Prade (ed.) 13th European Conf. on Artificial Intelligence, John Wiley & Sons, 1998.

Fayyad, U.; Piatetsky-Shapiro, G., Smith, P. and Uthurusamy, R. (eds.) *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, Cambridge Mass., 1996.

Fayyad, U.; Piatetsky-Shapiro, G., and Smith, P., From data mining to knowledge discovery: An overview, in (Fayyad et al. 1996a), 1996.

Fayyad, U.; Piatetsky-Shapiro, G., and Smith, P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, Vol. 39, No. 11, 27-34, November 1996.

Fox, C., Levitin, A., Redman, T., The notion of data and its quality dimensions, *Information processing and management*, 30(1), 1994.

Hernández-Orallo, J., Constructive Reinforcement Learning, *Intl. J. of Intelligent Systems*, to appear, 1999. URL: http://www.dsic.upv.es/~jorallo/escritsa/ IJISHern.ps.gz

Liepins, G. & Uppuluri, V. (eds.) *Data Quality control: theory and pragmatics*. M. Dekker, 1990.

Orr, K., Data Quality and Systems, *Communications of the ACM*, Vol. 41, No. 2, Feb. 1998.

Parsaye, P. & Chignell, M. *Intelligent Database Tools and Applications*, John Wiley & Sons, Inc., 1993.

Rakov, I., *Data Quality and its Use for Reconciling Inconsistencies in Multidatabase Environment*, Ph.D. Dissertation, George Mason University, Fairfax, Virginia, 1998.

Redman, T.C., Data Quality for the Information Age, Artech House, Boston, MA, 1996.

Rothenberg, J. "Metadata to support data quality and longevity" in *Proc. of the 1st IEEE Metadata Conference*, Silver Spring, MD, 1996.

Schlimmer, J. "Learning meta knowledge for database checking" in Proc. 9th Natl. Conf. on AI, 1991.

Wang, R.; Kon, H.; Madnick, S., Data Quality requirements analysis and modeling, in *Proc. of the 9th International Conference on Data Engineering*, 1993.

Wang, R.; Reddy, M.; Kon, H.; Toward quality data: An attribute-based approach, *Decision Support Systems*, 13(3-4), 1995.

Weiss, S. M.; Indurkhya, N. "Predictive Data-Mining: A Practical Guide" Morgan Kaufmann Pubs., 1997.