KNOWLEDGE ENGINEERING: PRINCIPLES AND TECHNIQUES
Proceedings of the International Conference on Knowledge Engineering,
Principles and Techniques, KEPT2007
Cluj-Napoca (Romania), June 6–8, 2007, pp. 282–289

DATA VERIFICATION IN ETL PROCESSES

MARIAN BALTA[1]

Abstract. The ETL processes are responsible for the extraction of the data from the external sources, transforming the data in order to satisfy the integration needs and for loading the data into the data warehouse. On the other hand, in the data mining world, there is a special concern on using the metrics for efficient classification algorithms. One of these approaches is the one that uses metrics on partitions based on the Shannon entropy (or other forms of entropy), to study the degree of concentration of values. In this paper we show how this idea can be used in verification of the consistency of data loaded into the data warehouse by ETL processes. We calculate the Shannon entropy and Gini index on partitions induced by attribute sets and we show that these values can be used to signal a possible problem in the data extraction process.

1. Introduction

In a data warehouse, the periodical integration of the data from the external sources trough the ETL processes produces large amounts of data. Even that each ETL process contains verification stages, it is possible that some particular kind of anomalies may occur and not be detected. These anomalies are not necessarily caused by a malfunction but it is very useful (if not mandatory) to find them. For example, if, at a given moment, one of the external data sources does not provide any data (or less data than the rest of the sources or than usual), then something may be wrong. The detection of such a scenario can easily be done using entropy defined on partitions. The idea was suggested to us by professor Simovici after a presentation he made on a summer school in Iasi [4].

Much work has been done in respect to the use of the entropies into the data mining field. Regarding the use of partition entropy you can find a complete formal description in [5]. Usually, in data mining, the notion of entropy is used to define metrics on partitions sets, in attempt to develop good classification algorithms. However, computing the entropy of a data set in respect to a given partition can

2000 *Mathematics Subject Classification.* 68P15, 68P20.
Key words and phrases. Data warehouse, ETL, Shannon entropy.

provide some very useful information regarding the distribution of values over the partition blocks.

The use of metrics in ETL field is not new. In [7], the authors model an ETL scenario as a graph and introduce specific importance metrics. In other works, like [2], the metrics are defined and used for evaluation of real industry ETL products. We use the Shannon entropy, calculated for a partition induced on a relational table by a set of attributes, to verify the consistency of data extracted by the ETL processes from the external sources. We also show that the choice of the attributes set plays a very important role in the efficiency of the method.

The paper is structured as follow. The sections 2 and 3 provide a short introduction in the definition of the Shannon entropy, Gini index and the partitions induced by attributes sets. The section 4 briefly presents the ETL notions, the dimensional model and an example used in section 5 to demonstrate how the entropy can be used to verify the consistency of the data. The paper ends with some concluding remarks.

2. Shannon entropy for partitions

The concept of entropy in information theory describes how much randomness there is in a signal or random event. It is formally defined by Claude E. Shannon in 1948 in his paper "A Mathematical Theory of Communication" [3]. In terms of a discrete random event X, with n possible states, the Shannon entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \; log \, p_i$$

where p_i is the probability of occurrence of the state i. We call H the entropy of the set of probabilities $p_1, \ldots, p_n$.

The quantity H has a number of interesting properties:

(1) H = 0 if and only if all the p_i but one are zero, this one having the value unity. Intuitively, this means that we know the outcome since only one event can occur. In any other situation, the value of H is positive.
(2) For a given n, H is a maximum and equal to $\log n$ when all p_i pi are equal (i.e., $\frac{1}{n}$). This is also intuitively the most uncertain situation.
(3) Suppose there are two events, X and Y, in question with m possibilities for the first and n for the second. Let *p(i,j)* be the probability of the joint occurrence of i for the first and j for the second. The entropy of the joint event is

$$H(X,Y) = -\sum_{i,j} p(i,j) \; \log p(i,j)$$

while

$$H(X) = -\sum_{i,j} p(i,j)\ \log \sum_{j} p(i,j)$$

$$H(Y) = -\sum_{i,j} p(i,j)\ \log \sum_{i} p(i,j).$$

It is easily shown that

$$H(X,Y) \leq H(X) + H(Y)$$

with equality only if the events are independent (i.e., *p(i,j)=p(i)p(j)*).The uncertainty of a joint event is less than or equal to the sum of the individual uncertainties.

(4) Any change toward equalization of the probabilities $p_1, p_2, \ldots, p_n$ increases H.

These properties qualify H as a measure of the uncertainty of the outcome of the event X.

To extend this to a partition, we have to observe that giving $\pi = \{B_1, \ldots, B_n\}$, a partition on a finite and nonempty set A, we can associate a random variable as follow:

$$X_\pi = \begin{pmatrix} {}^{in B_1} & & {}^{in B_n} \\ \frac{|B_1|}{|A|} & \cdots & \frac{|B_n|}{|A|} \end{pmatrix}$$

The Shannon entropy of π is defined as the Shannon entropy of X_π. This entropy can be used to measure the concentration of values in the partition. For example, if we have a set of 15 elements in A, and a partition having 5 blocks then, for each following cases we have:

- $|B_1| = 11, |B_2| = 1, |B_3| = 1, |B_4| = 1, |B_5| = 1 \Rightarrow H(X_\pi) = 1.3699$
- $|B_1| = 6, |B_2| = 6, |B_3| = 1, |B_4| = 1, |B_5| = 1 \Rightarrow H(X_\pi) = 1.8389$
- $|B_1| = 5, |B_2| = 4, |B_3| = 4, |B_4| = 1, |B_5| = 1 \Rightarrow H(X_\pi) = 2.0662$
- $|B_1| = 3, |B_2| = 3, |B_3| = 3, |B_4| = 3, |B_5| = 3 \Rightarrow H(X_\pi) = 2.3219$

It is known that the value of the Shannon entropy is proportional with the degree at witch the element are equally scattered among the blocks of the partition. The larger the entropy, the more the elements of A are scattered among the blocks of π.

Another way to have a measure of the distribution of the elements between blocks is to calculate Gini's index using the following formula [5]:

$$H_1(X) = 1 - \sum_{i=1}^{n} p_i^2$$

Using the same example from above we obtain the following values:

- $|B_1| = 11, |B_2| = 1, |B_3| = 1, |B_4| = 1, |B_5| = 1 \Rightarrow H_1(X_\pi) = 0.4444$
- $|B_1| = 6, |B_2| = 6, |B_3| = 1, |B_4| = 1, |B_5| = 1 \Rightarrow H_1(X_\pi) = 0.6666$

- $|B_1| = 5, |B_2| = 4, |B_3| = 4, |B_4| = 1, |B_5| = 1 \Rightarrow H_1(X_\pi) = 0.7377$
- $|B_1| = 3, |B_2| = 3, |B_3| = 3, |B_4| = 3, |B_5| = 3 \Rightarrow H_1(X_\pi) = 0.8$

In fact, both formulas are particular cases of the generalized entropy of partitions introduced by Daróczy in the following form:

$$H_\beta(\pi) = \frac{1}{1 - 2^{1-\beta}} \left(1 - \sum_{i=1}^{n} \left(\frac{|B_i|}{|A|} \right)^\beta \right)$$

It is easy to see that for $\beta = 2$ we obtain the Gini index and that $\lim_{\beta \to 1} H_\beta(\pi)$ is Shannon's entropy.

3. Partitions induced by attribute sets

A table T in a relational database is a set ρ of tuples on a set of attributes $A = \{A_1, \ldots, A_n\}$, where each tuple $t \in Dom(A_1) \times \ldots \times Dom(A_n)$. The set ρ is called the content of the table T [6]. Any subset of attributes $K \subseteq A$ induces a partition on the content of the table, denoted by π_K. For each different set of equal values for the projection of the table T on K, we have a corresponding block of the induced partition (Table 1). Some interesting results have been obtained

	$\cdots$	$\leftarrow K \rightarrow$	$\cdots$
t_1	$\cdots$	k_1	$\cdots$
t_2	$\cdots$	k_1	$\cdots$
t_3	$\cdots$	k_1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_h	$\cdots$	k_p	$\cdots$
t_{h+1}	$\cdots$	k_p	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_m	$\cdots$	k_r	$\cdots$
t_{m+1}	$\cdots$	k_r	$\cdots$

Table 1. The partition induced on a table T by the set of attributes K

in data mining field, regarding the classification algorithms using decision trees, using this induced partition [5].

4. ETL and the Dimensional Model

Extract, transform and load (ETL) is a set of processes that include, as the most important parts, the following:

(1) the identification of relevant information in source sistems;
(2) the extraction of that information;

(3) the integration of the information coming from multiple sources into a common format;
(4) the cleaning of the resulting data set, on the basis of database and business rules;
(5) the propagation of the data to the data warehouse.

Despite the popularity of relational normal forms, the data warehouse field has some particularities that need to be considered. This is the reason that has lead to the use of a new conceptual design model - the dimensional model. The dimensional modeling is a technique used in conceptual modeling, and its aim is to present the data in a standardized manner and to allow a very fast access, in order to support analytical processing. The model is, obviously, dimensional and it uses the relational model with some very important restrictions. Each dimensional model is composed from a table with a multiple key, which is called the fact table, and a set of (smaller) tables, called dimensions. Every one of the dimensions has a singular key, usually corresponding to one of the components of the fact table key.

A fact table, because it has a multiple key formed by two or more foreign keys, always represents a many-to-many relation. Another very important aspect of the dimensional model is that the fact table contains one or many numerical columns called measures, attached to the combination of keys which defines each row. An important property of these columns is that one has to be able to aggregate them. The importance reside from the fact that the applications that use this model almost never use a single record; usually they select hundreds, thousands or even millions of rows and submit them to an aggregation.

We can find very good examples of dimensional model in [1] and we will use a version of that example in the following. We consider a data warehouse containing information about sales (Figure 1).

The fact table has a multiple key formed by the *time_key*, *product_key*, *store_key* and *customer_key* attributes. These fields are foreign keys that define the relations between the fact table and the dimensions. The main purpose of this model is to allow fast aggregation of sales over every dimension.

5. Verifying consistency

Every notion mentioned above is well known and used in many areas. In the following section we give a description of the way Shannon entropy and Gini's index can be used to verify the consistency of the data that is extracted from the external sources.

Depending on the nature of the dimensions involved in the schema, we can identify important cases in which the data that enter the data warehouse is distorted by anomalies caused by the inaccessibility of some parts of data. As the time and

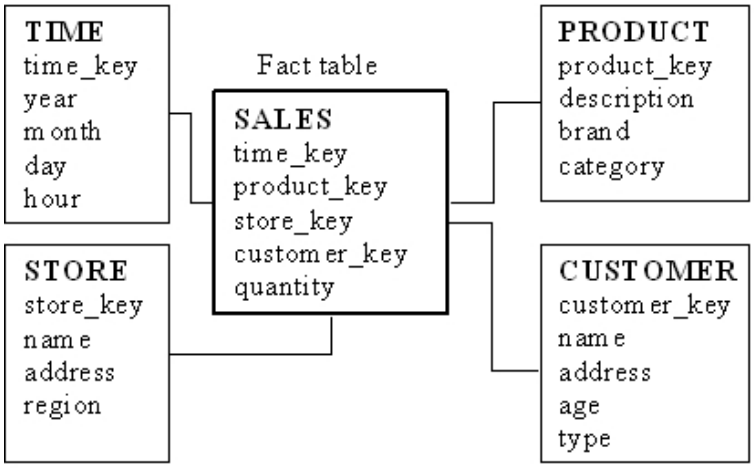

FIGURE 1. The dimensional model

spatial dimensions are always present in a data warehouse, let's consider one of the following situations:

(1) due to technical problems, one of the stores (or one of the regions) was unable to send the necessary information for data warehouse update, or,
(2) the missing data is from a specific period of time (let's say a day).

In the firs case we are interested in detecting if the sales from a store are missing. It is easy to see that we can define the partition induced on the table SALES by the attribute *store_key*. Let's assume that the sales are divided among the five stores, like in Table 2, during three consecutive days.

store / day	**1**	**2**	**3**	**4**	**5**
1	100,000	110,000	100,005	100,003	120,000
2	130,000	90,000	0	101,000	115,000
3	100,000	110,000	20,000	100,000	120,000

TABLE 2. The distribution of the sales among the stores

If we calculate Shannon entropy and Gini index for the induced partition over these three days, we have:

The results show that, if we carefully choose a level, we can signal as a possible fault any input for which the entropy falls bellow that level. However, the differences in values can have other cause than a technical problem and further testing is in order. For example, in the second day, from the 3^{rd} store there are no records.

day	H	H_1
1	2.3179	0.7989
2	1.9864	0.7453
3	2.1694	0.7684

TABLE 3. The Shannon entropy and Gini index using *store* dimension

We can see that this has a significant impact on the value of the Shannon: it has a smaller value than on the first day. This shortcoming in data can be caused by a technical problem but can easily be a normal situation (an event that required the closing of the second store that day). This is why the verification using the entropy must be followed by a further analysis.

For the second situation mentioned, in which the missing data is from a specific period of time, the counting of the sales has to be done for each store and analyzed in the same manner. The incoming data should be almost evenly distributed in time. So, this time we calculate the Shannon entropy and Gini index for each store, considering the partition induced by the attribute *time_key*. We have:

store	H	H_1
1	1.5734	0.6612
2	1.5788	0.6639
3	0.6502	0.2778
4	1.5849	0.6667
5	1.5847	0.6665

TABLE 4. The Shannon entropy and Gini index using *time* dimension

It is easy to observe that, in this case, the differences between the values of the entropy and Gini index for a regular store and for the store that has problems (3^{rd} store) are greater than in the previous case. This is due to the fact that for each store the values are more equally distributed among each day than they were among stores on a particular day.

From these two examples we can see that the choice of attributes used to define the partition have a great impact on the outcome. The attributes must be selected to assure that, in a normal situation, the number of elements (records) in every block of the induced partition is nearly equal.

6. Conclusions and further work

The notion of entropy and its applications in information theory are very powerful tools. The data mining field is using powerful algorithms obtained using this theory. By proper defining the involved parameters, it can be efficiently used for verification of the data loaded into the data warehouse.

In our approach, the value of the entropy remains high as long as the data from the operational sources is loaded into the warehouse on the regular basis. A disruption in this rhythm causes a change in the value of the entropy. However, we belive that this change can be magnified using some proper prior transformations of the values involved (subject to further work).

The generalization of partition entropy described in [4] could also be an interesting idea of study from the perspective of data warehouse environment.

References

[1] Kimball, R., "Drawing the Line between Dimensional Modeling and ER Modeling Techniques", http://www.dbmsmag.com/9708d15.html, 1997.
[2] Russom, P., Moore C., Teubner, C., "How To Evaluate Enterprise ETL", Forrester Research, 2004.
[3] Shannon, C. E., "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol. 27, pp. 379-423, 623-656, July, October, 1948
[4] Simovici, D., "Metric Methods in Data Mining", IDA 2006, Iasi, Romania, June 16, 2006.
[5] Simovici, D., Jaroszewicz, S., "An Axiomatization of Partition Entropy", Transactions on Information Theory, July 2002, vol. 48 (7), pp. 2138–2142.
[6] Ullman, J. D., "Principles of database systems", Computer Science Press, 1980.
[7] Vassiliadis, P., Simitsis, A., Skiadopoulos, S., "Modeling ETL activities as graphs", 4th Intl. Workshop DMDW'2002, Toronto, Canada, May 27, 2002, pp. 52-61.

(1) Computer Science Faculty, "Alexandru Ioan Cuza" University Iasi, General Berthelot, 16, 700483 Iasi, Romania

E-mail address: mbalta@infoiasi.ro