Storage and Bayesian modeling of data on the social resilience: Case of Orphans and Vulnerable Children (OVCs) in Côte d'Ivoire

Kouassi Bernard SAHA[1], Odilon Yapo M. ACHIEPO[2], Konan Marcelin BROU[3], Oumtanaga Souleymane[4]

[1]Laboratory of Computer Science and Telecommunications, National Polytechnic Institute Abidjan, 08 BP 475, Côte d'Ivoire

[2]Laboratory of Mathematics and New Information Technologies, National Polytechnic Institute Yamoussoukro, BP 1093, Côte d'Ivoire

[3]Department of Mathematics and Computer Science, National Polytechnic Institute Yamoussoukro, BP 1093, Côte d'Ivoire

[4]Laboratory of Computer Science and Telecommunications, National Polytechnic Institute Abidjan, 08 BP 475, Côte d'Ivoire

Abstract

A data storage system is a set of technologies and tools helping in decision making and storing optimal data for analysis. The data used for building data warehouses usually come from different sources and their processing mainly aims at improving the decision making process. The use of data warehouse allows a considerable gain in time spent on the development of requests for the search of decision making related-evidence induced by an increasingly competitive business environment. The purpose of this paper is to propose a dimensional model adapted to social resilience process and a Bayesian simulation model of OVCs' resilience mechanism due to HIV AIDS in Côte d'Ivoire. This work falls within the framework of the development of the Resilience Engineering which actually focuses on resilience measurement and analysis by computational modeling approaches. The significant modeling branch is the Resilometrics, creates by Achiepo Odilon Yapo [1] and [2] that is actively under development. The paper also intends to propose mathematical and computational tools to facilitate the work of the observatory of resilience in general and social resilience in particular. The use of orphans and vulnerable children context as a result of HIV-AIDS is an example of the actual implementation of the Resilience Engineering.

Keywords: *Data Warehousing, Dimensional Model, Resilometrics, Resilience Engineering, Bayesian Networks, Orphans and Vulnerable Children, HIV-AIDS.*

1. Introduction

In the original context of Business Intelligence, data warehouses were needed to better use the data. In a capitalist world governed by the principle of free competition and the profit motive, the development of assistive technology decision is a necessity with the IT Business Intelligence technologies. These are based on the concept of data warehouse, a basic conventional data storage approach, provided data models not using the classical entity-relationship formalism. The most used model to build data warehouses is the start schema. As part of the activities on resilience, the establishment of an observatory requires resilience to have data storage devices for analysis purpose. In this context, the use of data warehousing is an ideal choice. The data stored in the data warehouse may be used in Data Mining studies to optimize decision making. This article focus on the particular case of resilience analysis consisted in proposing a suitable approach dedicated to data warehousing and modeling the OVCs' resilience data in Côte d'Ivoire. Precisely, we use the Bayesian networks technology for the OVCs' resilience analysis in order to facilitate the understanding of the process governing OVCs' resilience in Côte d'Ivoire. The choice of the Bayesian network technology is due to the fact that these kinds of model have qualitative and quantitative aspects. This specificity make the model easily comprehensive for people not specialized in data modeling and can be used by many people without technical knowledge.

2. Dimensional modeling of OVCs resilience data

In Côte d'Ivoire the issue of support for orphans and vulnerable children (OVCs) has led many organizations to implement care arrangements monitoring and evaluation systems. Very often management policies are provided by international programs. The following table shows the different dimensions included in the monitoring and evaluation mechanism for the NGO Manasseh OVCs in Côte d'Ivoire.

Table 1: dimensions and monitoring and evaluation followed by OVCs

DIMENSIONS	ATTRIBUTES	LEVELS
Food and nutrition	Food Safety	Good, Average, Poor, very Bad.
	Growth and Nutrition	Good, Average, Poor, very Bad.
Housing and Care	Housing	Good, Average, Poor, very Bad.
	Care	Good, Average, Poor, very Bad.
Protection	Abuse and Exploitation	Good, Average, Poor, very Bad.
	Legal Protection	Good, Average, Poor, very Bad.
health	health	Good, Average, Poor, very Bad.
	Health Services	Good, Average, Poor, very Bad.
Psychosocial	Emotion	Good, Average, Poor, very Bad.
	Social behavior	Good, Average, Poor, very Bad.
Education and performance	Education	Good, Average, Poor, very Bad.
	Performance	**Good, Average, Poor,** very Bad.

The table shows the structure of monitoring and evaluation information to assess the level of vulnerability and resilience of OVCs proposed for the CSI (*Children Status record Index*). Based on this, one can construct an improved dimensional model, adapted to the setting up of a OVCs' monitoring and evaluation data warehouse. In fact, dimensional modeling including the star schema is well known for its effectiveness in developing solutions to aid the decision making. It is readily usable for the development of reporting applications and dashboards. The star schema from the Table 1 is given by the following diagram:

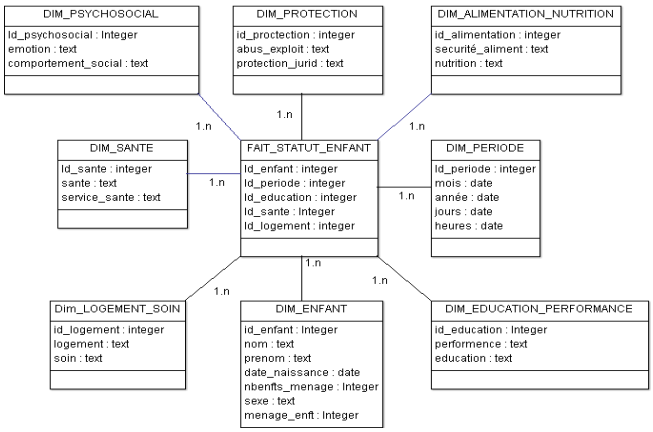

Fig 1: OVCs' resilience model in the star shaped diagram

The implementation of this diagram requires that the diagram database currently used (spreadsheet) be converted into the format of the data warehouse by an ETL program. On top of that, the quality of stored data should be considered in the practical implementation of the data warehouse. For high quality data, it is still possible to extract useful data for a multidimensional search, realizing, at the same time, all the data cleaning operations necessary. The advantage of having a data warehouse is to regularly study OVCs' resilience in order to optimize the management policies used. According to Richardson, "Resilience is the process of adaptation to stressors, adversity, changes and opportunities, resulting in the identification, strengthening and enhancement of protective factors, whether personal or environmental" (Richardson, 2002). Many non-governmental organizations for health and HIV-AIDS fighting are overwhelmed with data but lack the information needed in making good decisions. Knowledge Discovery in Database (KDD) technologies can help these organizations optimize their decision making. Therefore, in research and pratical works on resilience, the measurement and modeling technics are introduced, especially in the fields of Resilometrics, in order to adapt some modeling methods to resilience processes specificities. In the particular case of OVCs, analytical needs require easy way to update and providing a simple and efficient simulation approach. In this context, the Bayesian Networks technology is an ideal approach.

3. Bayesian network OVCs' resilience modeling

Also known as probabilistic expert systems, Bayesian Networks are tools of knowledge representation and automated reasoning on that knowledge. They were introduced by Judea Pearl in the 1980s and are found to be powerful useful tools for representing uncertain knowledge and reasoning from incomplete information. Bayesian Networks are simulation tools for observing the behavior of a complex system in contexts and conditions that are not necessarily accessible to experimentation. Technically, Bayesian networks are graphical models combining graph theory and probability theory. The following diagram (2) shows an example of Bayesian Network:

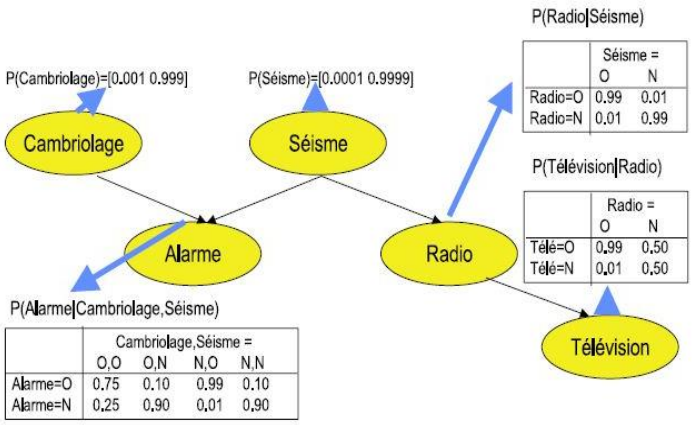

Fig 2: Example of Bayesian Networks

The above diagram is taken from a tutorial presented at the 8th scientific meeting dedicated to Knowledge Discovery in Data from (Philippe Leray). This Bayesian Network [11] and [12] models the process of triggering a security alarm in an environment frequently subjected to earthquakes. These earthquakes affect radio facilities upon which the television infrastructure is built. As shown in figure (Fig 2), a Bayesian network is a directed graph in which nodes represent the variables and the arcs symbolize the dependency relationships between these variables. Each node has a conditional probabilities table that is a model of beliefs in the occurrence of a particular case when we are in such a condition. In the case of modeling the process of identifying the best actions of resilience, such a graph translates the identified actions and decision variables thereon. This graph depends on the policies considered and the structure of the corresponding interactions that may vary from one study to another. To modeling the process of understanding the resilience of OVCs, let consider:

- $X = (X_i)_{1 \leqslant i \leqslant N}$ the different attributes considered in the CSI monitoring and evaluation system, all combined dimensions (nodes of relationships graph).
- K_i the number of levels of the attribute (node) X_i
- $\mathcal{P}(X_i)$ the group of variables that are parents nodes of the attribute node X_i

Attributes X_i (nodes) are linked by causal relationships. The following chart provides the structure of relationships between the different attributes:

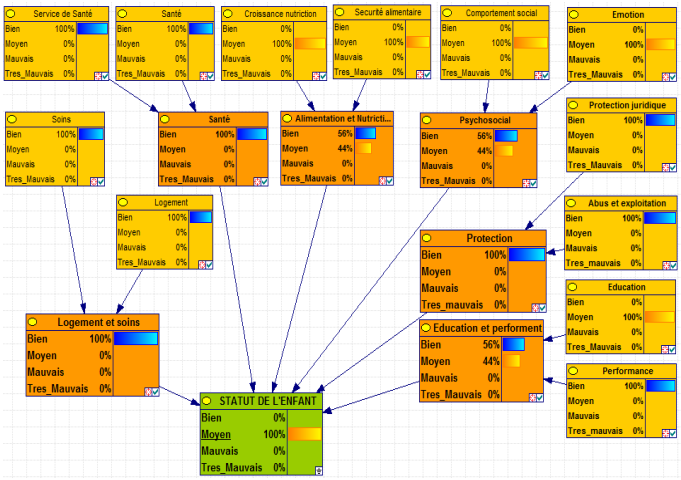

Fig 3: Structure of Bayesian Network

This graphical structure is an intuitive representation of the resilience process governing OVCs based on the information (attributes) used by the CSI to assess their related resilience. This dependency graph is the qualitative part (or aspect) of the corresponding Bayesian network model.

Formally, a Bayesian network is a couple (G, Θ) where:

- G is a directed graph without cycle
- Θ is a probability distribution defined on the variables of the graph (attached to the nodes)
- Each node of G is associated with a random variable and only one (an attribute X_i)
- The set $\{X_1, \ldots, X_N\}$ represent all random variables (nodes of the graph).
- The join probabilities of K nodes obeys the fundamental property below:

$$\Theta = \mathbb{P}(X_1, \ \ldots \ , X_K) = \prod_{i=1}^{K} \mathbb{P}(X_i | \mathcal{P}(X_i)) \qquad (1)$$

A Bayesian network is completely described when we have some nodes, the relationships graphs between nodes and the conditional probabilities associate to each node knowing each related parent. All Bayesian Network follows the Markov condition, that is to say, in a Bayesian Network, any node is conditionally independent of its descendants, knowing his parents. In the case of OVCs, inference in Bayesian networks is to calculate the probabilities:

$$\Theta = \mathbb{P}(X_1, \ \ldots \ , X_K) = \prod_{i=1}^{K} \mathbb{P}(X_i | \mathcal{P}(X_i))$$

In practice, these calculations are performed through chaining of rule following conditional probabilities $\forall K \in [1, N]$:

$$\mathbb{P}(X_1, \ \ldots \ , X_K) = \prod_{i=1}^{K} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \qquad (2)$$

4. Proof of the chain rules

The rule chain is not a general probability theory mechanism. It's important to factorization to reduce probabilities calculus in Bayesian network technology. This property can be proof easily. Using Bayes theorem, $\forall X_i, X_j$, we can write :

$$\mathbb{P}(X_i, X_j) = \mathbb{P}(X_i | X_j)\mathbb{P}(X_j) = \mathbb{P}(X_j | X_i)\mathbb{P}(X_i) \quad (3)$$

If $K = 2$, we have:

$$\mathbb{P}(X_1, X_2) = \mathbb{P}(X_2 | X_1)\mathbb{P}(X_1)$$
$$= \prod_{i=1}^{2} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \qquad (4)$$

If $K = 3$, we have:

$$\mathbb{P}(X_1, X_2, X_3) = \mathbb{P}(X_3 | X_2, X_1)\mathbb{P}(X_2, X_1)$$
$$= \mathbb{P}(X_3 | X_2, X_1))\mathbb{P}(X_2 | X_1)\mathbb{P}(X_1)$$
$$= \prod_{i=1}^{3} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \qquad (5)$$

Let assume that $\forall K \in [1, N]$:

$$\mathbb{P}(X_1, \ldots, X_K) = \prod_{i=1}^{K} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \quad (6)$$

Based on this assumption, we can write:

$$\mathbb{P}(X_1, \ldots, X_{K+1}) = \mathbb{P}(\{X_1, \ldots, X_K\}, X_{K+1})$$
$$= \mathbb{P}(X_{K+1}|X_K, \ldots, X_1)\mathbb{P}(X_1, \ldots, X_K)$$
$$= \mathbb{P}(X_{K+1}|X_K, \ldots, X_1) \prod_{i=1}^{K} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1)$$
$$= \prod_{i=1}^{K+1} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \quad (7)$$

Therefore, we have $\forall K \in [1 : N]$:

$$\mathbb{P}(X_1, \ldots, X_K) = \prod_{i=1}^{K} \mathbb{P}(X_i \mid X_{i-1}, \ldots, X_1) \quad (8)$$

Although the conditional probabilities can be provided by OVCs experts, the fact of having stored data provides more accurate estimates of these probabilities. We therefore perform parameter learning from data. Learning Bayesian Networks Parameters from data consists on estimate the unknown parameter based on the data. Knowing the structure, the data is assumed representing the probability distributions governed the studied process. To estimate the probabilities, if we note x_i a possibility we can observed for the attribute X_i, the likelihood function defined by:

$$\mathcal{L}(\Theta, \mathcal{D}) = \mathbb{P}(\mathcal{D}|\Theta) = \prod_{i=1}^{M} \mathbb{P}(x_i \mid \Theta) \quad (9)$$

The principle of learning the join probabilities Θ (parameters) using observed data consists on using the technic of maximum of the log-likehood defined by :

$$\mathbb{L}_{\mathcal{D}}(\Theta) = \log(\mathcal{L}(\Theta, \mathcal{D})) = \sum_{i=1}^{M} \log(\mathbb{P}(x_i \mid \Theta)) \quad (10)$$

In practice, because of the importance of Bayesian Networks, many algorithms have been developed to do these estimations and many software implements the different algorithms. The important thing with using Bayesian Networks consists in, once parameters are estimated, simulating the effects of a number of choices on all other actions and variables included in the model developed. The GeNIes Bayesian Networks software, used in this paper, is a very comprehensive software and especially free for both research work and for trade-related work. In fact, GeNIes is open source software dedicated to Bayesian Networks and their extensions (Dynamic Bayesian Networks and Influence Diagrams). It includes a large number of learning algorithms as well as the parameters of the structure from the data. It also has a friendly interface and can easily be used by non-specialists in modeling, including psychologists, economists, sociologists, environmentalists, etc.

5. Conclusions

As part of OVCs management policies, the storage of information collected in a data warehouse, not only will greatly improve the management of these data, but also their processing for purposes of decision support, particularly in understanding OVCs resilience processes. As part of the analysis of resilience in general and that of OVCs in particular, Bayesian networks are particularly appropriate because it is adapted to situations where one is faced with the uncertainty, incompletely and inaccurately. The use of the Bayesian simulation in the management of OVCs will allow organizations using the CSI, to better understand the process of resilience by simulating the impact of changes in one or more attributes of the CSI and can update easily the model with newly available data collected about the orphans and vulnerable children situation.

References

[1] Achiepo Odilon Yapo M, modélisation de la resilience; necessité d'une approche traitionnelle, 5e colloque international "Resilience en Action", Decembre 2014

[2] Achiepo Odilon Yapo M, Les bases fondamentales de la Résilométrie, une discipline de modélisation ded la souffrance. Journée scientifique "Café Resilience", Février 2015

[3] Cornuéjols Antoine Laurent Miclet, Apprentissage Artificiel, Concepts et algorithms, 2e edition, Eyrolles, 2010.

[4] Michallet Bernard, Ph.D., CRDP InterVal, GIRAFFE PIRC; 2nd annual conference of the CRDP InterVal Resilience and rehabilitation to follow a story. Dercon, S. 2001.

[5] FAO. The State of Food Insecurity in the World 2010: Adressing food insecurity in protracted crises. Rome, the United Nations Food and Agriculture and World Food Programme, 2010.

[6] G. Gardarin, "Internet and databases", Eyrolles, 1999.

[7] Martin-Breen, P. & Anderies, M. Resilience: A literature review. New York, USA, City University of New York and Tucson, USA, Arizona State University, 2011.

[8] R. Kimball, L. Reeves, M.Ross, W.Thornthwaite. "Designing and Deploying a Data Warehouse" Edition Eyrolles, 2000.

[9] R. Kimball, R. Mertz, "The Data Webhouse: Building the Web-enabled Data Warehouse", John Wiley & Sons, 2000.

[10] René Lefebure et al, Data mining, Eyrolles, 2001 392 pages.

[11] LERAY Philippe, Bayesian Networks: Learning and modeling of complex systems, HDR 2006.

[12] P.Naim, P. Leray et al, Bayesian Networks, 3rd edition, Eyrolle 2007.

[13] Jean-Michel Reinert, Les outils de la résilience, une force, un appui pour tous. Conférence-animation du 26 mars 2013, Chavannes-Renens, 2003.

Kouassi Bernard SAHA Is a Computer sciences Engineer (Agitel-Formation Abidjan, Côte d'Ivoire) and a Master degree holder in Computer Science with specialization in Industrial Computer sciences and Business Intelligence (University Nangui Abrogoua). He is a Ph-D student in Mathematics and Information Technologies (EDP INP-HB Yamoussoukro, Côte d'Ivoire). He is also a Teacher-researcher at the University Felix Houphouët Boigny (Côte d'Ivoire), and member of the Research Laboratory in Computer Sciences and Telecommunications of Houphouët Boigny National Polytechnic Institute (INP-HB), Abidjan, Côte d'Ivoire. His interests of the research include The Data Mining, and. his works are centered on their research of the database and programming languages.

Odilon Yapo M. ACHIEPO is a statistician-economist Engineer (ENSEA Abidjan, Côte d'Ivoire) and has a Master degree in Computer Science with specialization in Artificial Intelligence and Business Intelligence (Polytechnic School of Nantes, France). Ph-D student in Mathematics and Information Technologies (INPHB Yamoussoukro, Côte d'Ivoire). He is also an International Senior-Expert Consultant, member of the international resilience research group (UMI Resilience, IRD) and is a member of the Laboratory of Computer Sciences and Telecommunications (INP-HB) Abidjan, Côte d'Ivoire. His principals centers of interests are Computational Mathematics, Multidimensional Statistics, Artificial Intelligence, Multi-Agents Systems, Machine Learning, Data Mining, Data Science, Pattern Recognition, Embeded Systemes and Robotics. He also is the author-creator of the Resilometrics, a modeling discipline which consists on developing and applies computational models for measure, analyze and simulate social resilience process.

Konan Marcellin BROU is Doctor in Computer Science and Teacher researcher at the Houphouët Boigny National Polytechnic Institute (INP-HB) of Yamoussoukro (Côte d'Ivoire). He is the Director of the Department of Mathematics and Computer Science. He is a Member of Laboratory in Computer Sciences and Telecommunications (INPHB) Abidjan Côte d'Ivoire. His interests are information systems, database and programming languages.

.

Souleymane Oumtanaga is a Professor in Computer Science and Teacher researcher at the Houphouët Boigny National Polytechnic Institute (INP-HB) de Yamoussoukro (Côte d'Ivoire). He is the Director of the Laboratory in Computer Sciences and Telecommunications of Houphouët Boigny National Polytechnic Institute (INP-HB), Abidjan, Ivory Coast. His interests are Futur Network systems, Network security and Telecommunications.