Strategy for Generation of Knowledge through Automatic Correlation of Dimensional Data in Star Schema: Application in the Context of Leishmaniasis

Wallace Anacleto Pinheiro[1,2][0000-0001-7076-8785], Geraldo Xexéo[1][0000-0003-3975-9076], Jano Moreira de Souza[1][0000-0001-5080-1955], Ana Bárbara Sapienza Pinheiro[3][0000-0002-2766-3141], Ciro Gomes[3][0000-0002-3069-6884]

[1] COPPE – Graduate School and Research in Engineering – Universidade Federal do Rio de Janeiro (UFRJ) – Rio de Janeiro, Brazil
[2]IME – Instituto Militar de Engenharia – Rio de Janeiro, Brazil
UNB – Universidade de Brasília, Brasília, Brazil
wallaceapinheiro@gmail.com
xexeo@cos.ufrj.br
jano@cos.ufrj.br
absapienza@gmail.com
ciromgomes@gmail.com

Abstract. The aim of this article is to propose a methodology for generation of knowledge through correlation techniques, based on dimensional data obeying a star model. This methodology relates and integrates data from one or more thematic areas (data marts) using correlation coefficients applied to data derived from facts and dimensions of a data mart. In order to analyze the application of the strategy, we selected the database of Brazil Information System for Notifiable Diseases (SINAN) of Brazilian Department of Health to gather information about a particular neglected disease that affects several Brazilian capitals: Visceral Leishmaniasis. The proposed methodology has generated several correlations that can support the planning of public strategies to combat this disease.

Keywords: Correlation, Star Schema, Leishmaniasis, Neglected Diseases.

1 Introduction

Media in Brazil frequently disclose information on outbreaks of diseases in places that were not previously endemic [1]. Trade, means of transportation, foreigners without immunity to local pathologies, residents without immunities to foreign pathologies are elements that mix, narrow borders, break geographic boundaries, connect cities and diseases as well.

Data on climate, characteristics of the inhabitants of a region and characteristics of the place where they live can provide valuable information to understand and plan actions that seek to prevent or minimize the outbreak of diseases that spread among various cities in the world.

Thus, this article proposes the use of correlation techniques on data related to a neglected disease that affects several Brazilian cities and municipalities: Visceral Leishmaniasis. From 2007 to 2015, in Brazil, the annual average number of reported cases related to all types of leishmaniasis exceeded 20,000 notifications, of which 4,000 notifications reported visceral leishmaniasis. According to some authors [2], there is an increasing urbanization of some types of Leishmaniasis in certain Brazilian cities. In the case of a vector-borne parasite (ex.: sandflies), some indicators may be essential for the survival of the parasite and the vector considering the environment. The vector, in particular, requires the presence of some freshwater source (rivers, lakes, etc), having an area of action of only some hundreds of meters from these sources, due to its small size. It is a hematophobic insect, also known as sandfly. With this knowledge, some authors [3] report that increased temperature and rainfall precipitation contribute to the appearance of new cases of leishmaniasis. Other authors [4] believe that the dry climate and low altitudes contribute to the presence of the disease, since they benefit the presence of the vector.

Therefore, considering the increasing acting range of vector, and that the disease occurs in different clinical forms, from the tegumentary to the visceral, several cities in the American continent, especially some Brazilian cities, are vulnerable to this growing epidemiological problem [5]. Although there are different types of leishmaniasis, depending on the parasite and the immunity of the individual, the process of contagion is, basically, the same. Deforestation, climate change, the reservoir with several hosts, the peri-urban area, ecological tourism, among others, are some risk factors. It is estimated that there are two million new cases of leishmaniasis in the world each year [6], mainly in third world countries.

This work proposes that the notification data on Visceral Leishmaniasis stored in Brazil Information System for Notifiable Diseases (SINAN) of Brazilian Department of Health be correlated, seeking to relate characteristics of the affected population with the increase or decrease in the number of cases. It also proposes to correlate data from SINAN with data on climate provided by Brazilian Institute of Research and Forest Studies (IPEF). Thus, we intend to obtain an overview of possible links between factors that may contribute to the increase or reduction of the number of cases of leishmaniasis according to each region.

In this work, we do not discuss the use of causality techniques, we only analyze correlations between data and their coherence.

The rest of this work is organized as follows: Section 2 presents the related works, Section 3 presents the experiments configuration, Section 4 presents results and Section 5 shows the final considerations and the proposals for future work.

2 Related Work

SINAN is a system provided by Brazilian Department of Health that stores data on compulsory notification diseases or injuries in Brazil [7]. It provides historical data, allowing them to be analyzed on different aspects. Some works use SINAN data to relate diseases or injuries with climatic factors or characteristics of a region or work

[3], [8]. However, these studies use specific models to relate data in specific regions, not using correlation factors for a broader analysis.

In the area of information technology, some researchers [9] [10] cite data mining techniques that can be used in different types of databases, including data marts and data warehouses, such as trees Bayesian classification, association rules (correlation), neural networks, clustering, and genetic algorithms.

There are studies aimed at correlating datasets that represent multidimensional distributions using correlation coefficients, such as Pearson coefficient [11]. Correlation is done using characteristics of the distribution, such as linearity and non-linearity. These works are not related to dimensional modeling used in a data warehouse (dimension and fact), but addresses multi-dimensional data distributions. However, we did not find in literature works that perform correlation formed by combining a dimension of a star schema with time dimension from fact metrics.

3 Experiment Configuration

The approach proposed in this work can be applied in different areas of knowledge, since information be stored in a star schema based repository. Thus, we decide to apply it in the context of tropical neglected diseases and climate, a relevant topic, since the world is passing through many climate changes, while some tropical neglected diseases are spreading in new regions [2], [4], [5].

This work materializes the proposed strategy using data from SINAN of Brazilian Department of Health, available at: http://portalsinan.saude.gov.br/, and data from Brazilian Institute of Research and Forestry Studies (IPEF), available on the website: http://www.ipef.br/geodatabase/. Both SINAN and IPEF provide their data in file format. It was chosen to extract data on a specific neglected tropical disease: Visceral Leishmaniasis.

Leishmaniasis data, extracted from SINAN, were combined with climatic data on Brazilian municipalities, extracted from IPEF, in a single data repository with two data marts that served as basis for correlations.

Fig. 1 presents an overview of the process executed for creation of the dimensional database, and the element highlighted in gray corresponds to the components of the solution proposed in this work.

In this context, we executed the task of obtaining the data through an ETL process, using Pentaho Data Integration tool, Community version [12]. Extraction-Transformation-Load (ETL) tools, such as Pentaho, are widely used to integrate data from heterogeneous sources.

After passing through the ETL process, Data Correlation and Filtering module calculated the correlations. The implementation used Python 3.6 language and MicroStrategy Desktop 10.8.0 tool.

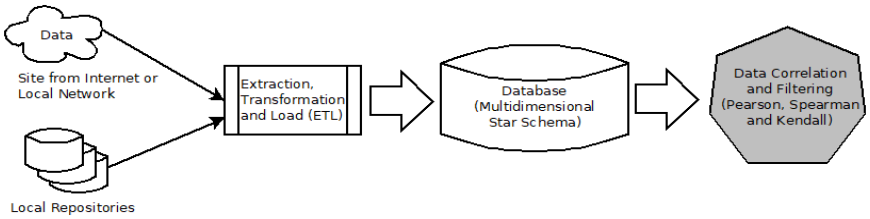

Fig. 1. – Proposed Architecture

Once stored in the database following the specification of a star schema, the proposed solution processed data in order to provide the correlations between the modeled dimensions. The number of processed data is showed in Table 1. It is possible to notice that few registers from the queries generate a large number of correlations.

Table 1. Quantity of Processed Data

Description of the Result	Quantity
Number of Registers involving Climate Fact (IPEF)	1500
Number of Registers involving Notification Fact (SINAN)	6590
Number of Correlations - Intra SQLs (Climate Fact)	1500
Number of Correlations - Intra SQLs (Notification Fact)	87740
Number of Correlations - Between SQLs (Climate Fact)	3750
Number of Correlations - Between SQLs (Notification Fact)	82496
Number of Correlations - Between Facts	73000

4 Results

Using the implementation discussed in previous sections, we have processed data stored on Leishmaniasis considering the period from January 2016 to December 2016 of Brazilian capitals. We considered only correlations greater than 50% (direct or inverse), having a significance level of at least 1%.

Fig. 2 shows the correlations found during this process. The label of the nodes identifies the correlated data, the size of the nodes as well as the width of the arcs are related to the number of correlations of the node, the color of the arcs represents the intensity of the correlation (the greater the intensity of the red, closer to the lower value of the correlation filter; while this, the greater the green intensity, the closer the upper value of the correlation value). Arrows indicate the direction of correlation. Names in the graph correspond to Brazilian capitals. Numbers ranges are related to age groups. Code *B550* corresponds to the International Code of Diseases (ICD) for Visceral Leishmaniasis (total cases). Some details of these correlations are discussed in the next paragraphs.

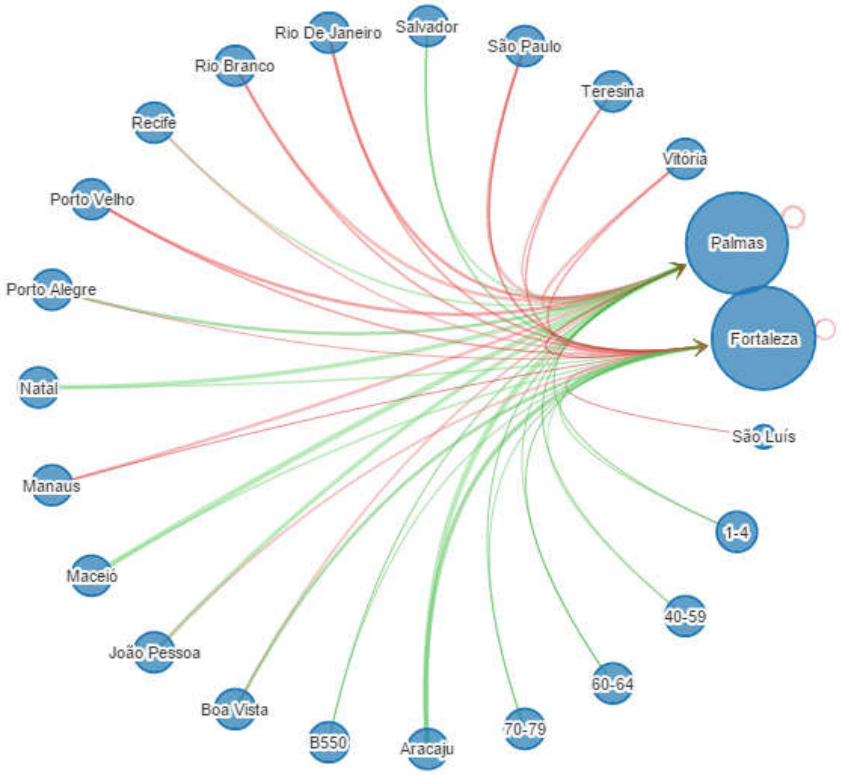

Fig. 2. – Correlations

Fig. 3, which details some correlations presented previously about temperature, rainfall index and number of cases among Brazilian capitals (these data are from two different data marts). Considering temperature, the highest value occurs between Palmas and Fortaleza (value of approximately 86%). When analyzing the rainfall indexes, the highest correlation values are obtained for Fortaleza and Aracaju (48%). Considering the relationship between the case number, it is observed that Fortaleza and Palmas are correlated (value slightly above 50%).

It is possible to observe that correlations 1 and correlations 3 contains Fortaleza and Palmas. These results suggest that, in the case of these two capitals, the temperature variation influences more the variation of the number of cases than the rainfall index.

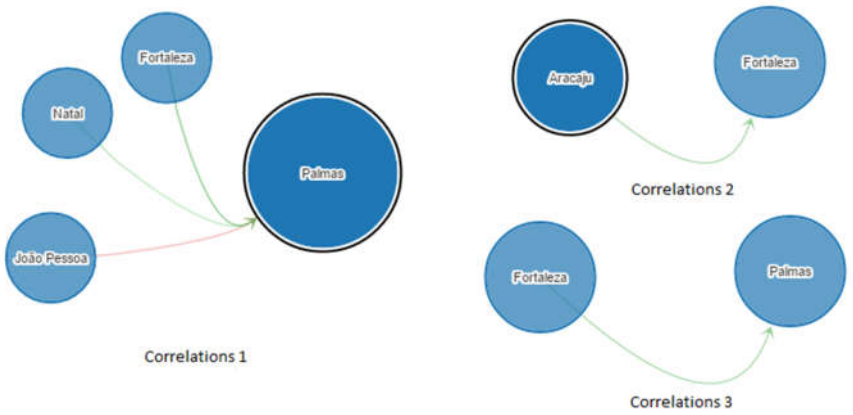

Fig. 3. Correlations for Temperature (Correlations 1), Correlations for RainFall Index (Correlations 2) and Correlations for Number of Cases (Correlations 3)

However, if we look at the other spectrum of correlation, that is, the inverse correlations, we can see that there is an inverse correlation higher than 70% between the number of cases in Palmas and the rainfall index of several capitals, including Palmas, observed in the self-relationship, as shown in Fig. 4.

This shows that in times of less rainfall, the number of cases in Palmas tends to increase. As seen earlier, higher temperatures also increase the number of cases in this region, providing a picture of the most critical periods in which vector-control measures should be taken. This, in a first moment, is not in agreement with the view presented by certain authors [3] that with increasing humidity there is an increase in the number of cases, but corroborates the view presented by others [4] who believe that Leishmaniasis is related to dry climates. In this scenario, the great advantage of the proposed strategy is to be able to base its observations on a large number of data from all notifications of Visceral Leishmaniases cases, considering Brazilian territory and not just local information.

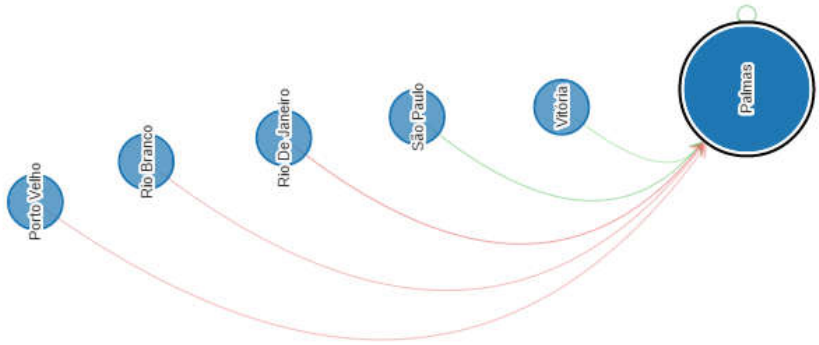

Fig. 4. – Inverse correlations considering rainfall index

Another possible analysis involves data from two different tables of the data mart for notifications, it is verified that there is a correlation higher than 50% between cases of leishmaniasis in Fortaleza for 60 to 64 age range, shown in Fig. 5. This can guide the planning of public policies to combat this disease in this place. Thus, it is possible to observe that, from a visual analysis of the correlation graph, several information can be obtained, providing information for potential assertive strategies by decision makers.

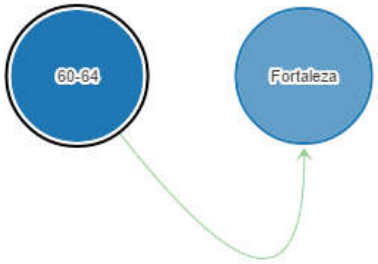

Fig. 5. Correlations between Fortaleza and Age Range

5 Final Considerations

Correlations between dimensions or their data can aid in generating new insights as they provide measurements of relationships over data marts that normally store a large amount of data.

The analysis of correlations of Brazilian capitals brought some relevant information: in some regions there was a direct relation between temperature and number of cases and an inverse relationship between rainfall index and number of cases. It is also possible to associate variation of the number of cases with specific age groups. These data could be used to guide public policies in Brazilian capitals, aiming to support the epidemiological control of leishmaniasis.

It is worth mentioning that previous studies on correlations involving Leishmaniasis have focused on specific sites, evaluating factors such as: climate (for example: rainfall and temperature) and population characteristics. However, a broader view that correlates a large volume of data provides a new light on this subject, since temperature, rainfall index and population characteristics are only some of the factors that may contribute to increase the number of cases of this disease.

Therefore, future studies will add new information from data sources considering other neglected diseases. The objective will be to analyze, in addition to correlations between the factors presented in this study, the effect of coinfections on evolution of diseases; in particular those transmitted by vectors.

References

1. Oliveira, T. Leishmaniose atinge mais municípios e avança em direção à capital (in portuguese). Veja - São Paulo (2018).
2. Bevilacqua, P. D., Paixão, H. H., Modena, C. M. & Castro, M. C. P. S. Urbanization of visceral leishmaniose in Belo Horizonte, Brazil. Arq. Bras. Med. Veterinária e Zootec. 53, 1–8 (2001).
3. Mendes, C. S., Coelho, A. B., Féres, J. G., Souza, E. C. de & Cunha, D. A. da. Impacto das mudanças climáticas sobre a leishmaniose no Brasil (in portuguese). Cien. Saude Colet. 21, 263–272 (2016).
4. Pinto, I. de S., Santos, C. B. dos, Grimaldi Jr., G., Ferreira, A. L. & Falqueto, A. American visceral leishmaniasis dissociated from Lutzomyia longipalpis (Diptera, Psychodidae) in the State of Espírito Santo, Brazil. Cad. Saude Publica 26, 365–372 (2010).
5. CDC. CDC - Leishmaniasis - Disease. Centers for Disease Control and Prevention (2013).
6. MS. Manual de vigilância da leishmaniose tegumentar (in portuguese). Ministério da Saúde (2017).
7. Barbosa, D. A. & Barbosa, A. M. F. Avaliação da completitude e consistência do banco de dados das hepatites virais no estado de Pernambuco, Brasil, no período de 2007 a 2010 (in portuguese). Epidemiol. e Serviços Saúde 22, 49–58 (2013).
8. Albuquerque, P. C. C. de et al. Health information systems and pesticide poisoning at Pernambuco. Rev. Bras. Epidemiol. 18, 666–678 (2015).
9. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. Advances in knowledge discovery and data mining. (AAAI Press, 1996).
10.Han, J., Kamber, M. & Pei, J. Data mining: concepts and techniques. (Elsevier Science, 2011).
11.Zhang, Y., Liu, T., Li, K. & Zhang, J. Improved visual correlation analysis for multidimensional data. J. Vis. Lang. Comput. 41, 121–132 (2017).
12. M., V., Syed, A., Mohammad, A. & N., M. Pentaho and Jaspersoft: A Comparative Study of Business Intelligence Open Source Tools Processing Big Data to Evaluate Performances. Int. J. Adv. Comput. Sci. Appl. 7, (2016).