Universidad del Cauca
Facultad de Ingeniería Electrónica y Telecomunicaciones
Programas de Maestría y Doctorado en Ingeniería Telemática
Seminario de Investigación
II - 2015

FRAMEWORK FOR DATA QUALITY IN KNOWLEDGE DISCOVERY TASKS (FDQ-KDT)

David Camilo Corrales Muñoz
Estudiante de Doctorado
13 de Noviembre de 2015

The presentation was composed for the next items:

1. BACKGROUND

In this section we presented the main definitions of data quality frameworks and the methodologies: Knowledge Discovery in Databases (KDD), Cross Industry Standard Process for Data Mining (CRISP-DM), Sample, Explore, Modify, Model and Assess (SEMMA) and Data Science.

- **Data Quality Frameworks (DQF):** DQF seek to assess areas where poor quality processes or inefficiencies may reduce the profitability of an organization [1]. At its most basic, a data quality framework is a tool for the assessment of data quality within an organization [2]. The framework can go beyond the individual elements of data quality assessment, becoming integrated within the processes of the organization. Eppler and Wittig [3] add that a framework should not only evaluate, but also provide a scheme to analyze and solve data quality problems by proactive management.

- **Knowledge Discovery Database (KDD):** Knowledge Discovery Database is defined as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [4]. The KDD process is interactive and iterative (with many decisions made by the user), involving numerous steps, summarized as: learning the application domain, creating a target dataset, data cleaning and preprocessing, data reduction and projection, choosing the function and algorithm of data mining, interpretation and using discovered knowledge [5].

- **Cross Industry Standard Process for Data Mining (CRISP-DM):** CRISP-DM is a comprehensive data mining methodology and process model that provides anyone from novices to data mining experts with a complete

blueprint for conducting a data mining project. CRISP-DM breaks down the life cycle of a data mining project into six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. [6].

- **Sample, Explore, Modify, Model and Assess (SEMMA):** The SEMMA process was developed by the SAS Institute that considers a cycle with 5 stages for the process: Sample, Explore, Modify, Model, and Assess. Beginning with a statistically representative sample of your data (sample), SEMMA intends to make it easy to apply exploratory statistical and visualization techniques (explore), select and transform the most significant predictive variables (modify), model the variables to predict outcomes (model), and finally confirm a model's accuracy (assess) [7].

- **Data Science:** The data science is focused in the representation, analysis, anomalies of data, and relations among variables, from a process with the next steps: raw data collected, data processing, clean data, exploratory data analysis, models and algorithms, construction of reports, and build data product [8]. Data Science employs techniques drawn from many fields within the broad areas of mathematics, statistics, signal processing, probability models, machine learning, statistical learning, data engineering, pattern recognition etc., with the aim to extract knowledge from data [9].

2. MOTIVATION

Data explosion is an inevitable trend as the world is connected more than ever. It is obvious that we are living a data deluge era, evidenced by the sheer volume of data from a variety of sources and its growing rate of generation. For instance, an International Data Corporation (IDC) report [1] predicts that, from 2005 to 2020, the global data volume will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, representing a double growth every two years [2]. The most fundamental challenge is to explore the large volumes of data and extract useful knowledge for future actions [3, 4]. For a successful process of discovery knowledge exist recognized methodologies such as Knowledge Discovery in Databases (KDD), Cross Industry Standard Process for Data Mining (CRISP-DM) and Sample, Explore, Modify, Model and Assess (SEMMA) which describe the data treatment. Similarly, the Data Science area searches the knowledge extraction with different approaches as stochastic modeling, probability models, signal processing, pattern recognition and learning, etc [5]. Although the knowledge discovery methodologies and data sciences defined the steps for data treatment, these not tackle the issues in data quality clearly, leaving out relevant activities [4]. It has been agreed that poor data quality in data mining, machine learning and data science will impact the quality of results of analyses and that it will therefore impact on decisions made on

the basis of these results. Different researchers have meanwhile shown the use of artificial intelligence algorithms to solve data quality issues in knowledge discovery tasks such as: heterogeneity, outliers, noise, inconsistency, incompleteness, amount of data, redundancy and timeliness [6], nevertheless heretofore there is not tools that integrate the algorithms for solve the data quality issues;
besides the data miners, data scientists, and anybody kind of related user do not know which is the the suitable algorithm for a data quality issue determined. Based on the considerations previously described, the present Doctoral project arises the next research question: ¿How assess the data quality in knowledge discovery tasks through artificial intelligence algorithms?

3. PROPOSAL

The proposed framework was developed to address poor quality data in knowledge discovery task such as data mining and machine learning projects. The process began with Gathering the preselected elements of CRISP-DM, SEMMA and Data science area. Afterward in Filtering & Mapping Phase the repeated components were removed. The Clustering phase, grouped the remaining components in five phases: data fusion, data quality diagnosis, select data, clean data, and construct data. Flow of development for FDQ-KDT The result to apply the methodology of Almutiry [10], is the FDQ-KDT which comprising seven tasks of the phases of data understanding and data preparation of CRISP-DM (describe data, explore data, verify data quality, select data, clean data, construct data, and integrate data), three stages of SEMMA (modify, sample, explore), and three steps of Data Science Area (clean data, exploratory data analysis and construction of reports), organized in five main phases. In this regard, the FDQ-KDT phases have an execution order with the aim of supply a tidy dataset.

4. CONCLUSION AND FUTURE RESEARCH

In knowledge discovery tasks such as classification, prediction, cluster, etc, is very important to use tidy dataset to get relevant outcomes. In the early decades of computing, a common saying was “garbage in, garbage out.” That is, mistakes in recollection of information were aberrations, and if knowledge discovery tasks have bad data (garbage in), then they should expect incorrect answers (garbage out) [11]. For this reason we proposed a conceptual framework for data quality in knowledge discovery task based on CRISP-DM, SEMMA and Data Science Area, which tackle the issues in data quality clearly through ESE taxonomy. Several approaches exist to tackle the issues of data quality in outliers, noise, inconsistency, incompleteness, and redundancy, amount of data, heterogeneity, and timeliness. Nevertheless the results to date not consider resolve the issues in ensemble. Thus the next step will be developing, examining and evaluating the proposed framework through artificial intelligence algorithms, statistical and mathematical models.

REFERENCES

[1] Karolyn Kerr and Tony Norris. The Development of a Healthcare Data Quality Framework and Strategy. In Proceedings of the Ninth International Conference on Information Quality, pages 218–233, 2004.

[2] Richard Y. Wang and Diane M. Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. J. Manage. Inf. Syst., 12(4):5–33, March 1996. 14

[3] Dörte Wittig. Conceptualizing Information Quality: A review of information quality frameworks from the last ten years. In Proceedings of the 2000 Conference on Information Quality, pages 83–96, 2000.

[4] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Advances in knowledge discovery and data mining. chapter From Data Mining to Knowledge Discovery: An Overview, pages 1– 34. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1996.

[5] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The kdd process for extracting useful knowledge from volumes of data. Commun. ACM, 39(11):27–34, November 1996.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. CRISP-DM 1.0 Step-by-step data mining guide, 1999.

[7] Delen Dursun Olson, David L. Advanced Data Mining Techniques. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[8] F. Pacheco, C. Rangel, J. Aguilar, M. Cerrada, and J. Altamiranda. Methodological framework for data processing based on the Data Science paradigm. In Computing Conference (CLEI), 2014 XL Latin American, pages 1–12, September 2014.

[9] Cathy O'Neil and Rachel Schutt. Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, 1 edition edition, November 2013.

[10] O. Almutiry, G. Wills, A. Alwabel, R. Crowder, and R. Walters, "Toward a framework for data quality in cloud-based health information system," in 2013 International Conference on Information Society (i-Society), 2013, pp. 153–157.

[11] A. Phalgune, C. Kissinger, M. Burnett, C. Cook, L. Beckwith, and J. R. Ruthruff, "Garbage in, garbage out? An empirical look at oracle mistakes by end-user programmers," in 2005 IEEE Symposium on Visual Languages and Human-Centric Computing, 2005, pp. 45–52.