PREPROCESSING STRATEGY IN WEB-MINING: RECOMMENDED OR INEVITABLE?

Dror Ben-Ami

Zefat Academic College, Zefat, Israel

ABSTRACT

Web-browsing users' behavior is one of the fascinating, attractive and interesting subjects, specially from socio-technological perspectives. Companies and learning-organizations are investing huge human resources, efforts and capital to follow their users' behavior, trying to find out and plot their users' profiles. Knowing, understanding and predicting users' tangible and intangible behavior help these organizations to focus on users' needs and interests. Thus, companies and learning-organizations can generate significant intellectual capital and use this asset efficiently and effectively later. The intellectual capital asset can target into business-based and knowledge-based activities, such as pointed-newsletters, targeted markets, crystalizing pricing policies and much more. The users' profiles usually categorized into different classifications, such as social aspects, personality traits, purchase behavior, cognitive behavior, and more. The raw data for this research was collected from web users. Around one hundred thousand internet-society users from one the OECD countries were the basis of the research population. The research examines users' behavior through the net, after collecting wide range of data elements during few months period, using progressive data analysis tools and techniques, afterwards. The case study relies of real work (project), conducted from July-2018 till Feb-2019. The purpose was to recover and re-analyze the collected data after performing improper and incorrect analysis procedures by the original staff. The paper deals with the possible missteps which can happen during the data preprocessing steps in data-intensive projects and tries to understand the consequences of those early missteps can have on the end result. Specifically, the paper recounts those missteps based on real experience during a web mining project. The paper presents a guide on the procedures and decisions which should be taken to avoid or at least minimize critical mistakes during the data preprocessing step.

KEYWORDS

Internet-society, Data/Web Mining, Preprocessing, Machine Learning, I2K (Information to Knowledge)

1. INTRODUCTION

The internet-society is a global worldwide community. It consists of internet users from different continents, different countries, different cultures and different browsing habits. This case study research was arranged in one European internet-society population; trying to explore and plot the following:

- a) Socio-technological behavior aspects of the internet-society.
- b) Follow communication and collaborations between internet users, from within a specific sub-group.
- c) Follow the population's relationships with the global worldwide internet-society community.

The case study is focused on the scientific data analysis continuous process, from its inter-organizational working process and procedures, and NOT focus in the Internet-Society itself. Specifically, the following pinpoints the sub *data-preprocessing* stage. A large data sets were assembled and monitored for few months during the research project. More than one hundred thousand internet users take part in this project. Fifteen to twenty different web browsing attributes and traits were collected, arranged and categorized, separately and within sub-groups categories. Ten research variables were the basis of the research data analysis process, after collecting, arranging and classifying the relevant data elements.

2. WEB MINING MODEL

The purpose of the researches' team was clearly defined: to explore and learn wide range of data behavior aspects, discover behavioral patterns and reinforce known paradigms and templates. The staff was interested to establish data-mining methods and techniques, rather than using the standard and conventional statistical analytical tools. The preference of data mining tools, as scientifically powerful analysis environment, rely on the facts that descriptive and predictive data mining models (Awad et al) (Dunham) can find out wide, deep, tangible and intangible data aspects, by using systematic machine learning rule-based, rather than those offered by the conventional data analysis tools, which are limited (Sivakumar et al) (Fauset). At first glance and based on this assumption, it seems clear and simple. Well-known and well-defined straight-forward process was approved; according to this scheme (Uwe).

Figure 1. Data Mining – Information to Knowledge Process

One of the main and central mile-stones in the DM general model is the preprocessing stage. Critical faults and significant mistakes were discovered and recognized in the *Preprocessing* stage. These, as expected, affected all the next data analysis stages. Wide range of references covers the issue of DM-preprocess stage/s from different points of view, but from one common significant angle: *the necessity to establish correct and proper step by step methodology, in order to increase the ability to analyze the data on the best way*; to achieve real valuable KNOWLEDGE, in the entire Information to Knowledge [I2K] continuous process. Kaur et al, refer in their article to effective pre-processing methodology, which involves field extraction, significant attributes selection, data selection, and data cleaning. Krell et al, focus on data selection strategies which limit the size of the stored training data by different inclusion, exclusion, and further dataset manipulation criteria. Similar reference can be found at Hatzi, et al. They emphasize that DM is long-continuous process, which sometimes might be cumbersome task, which requires expert knowledge, concerning preprocessing of data, selection of algorithms, appropriate configuration, feature identification and result comprehension. Two more such references are mentioned in the following articles and books. The first is specifically to the web-mining preprocess (Valesquez et al), and the second notes relevance for any data-mining analysis application, in general (Lawrence et al) (Michael et al). Feelders et al specify: "Despite the predominant attention on analysis, data selection and pre-processing are the most time-consuming activities, and have a substantial influence on ultimate success."

All the references express in different words, the required methodology and the accurate-proper preprocess "must-be-done" sub-activities. The project which relies on these articles' recommendations and inevitable steps, exposes few issues which cause to wrong process. Something went wrong along the socio-technological project. What was it? why? and could it be recovered and fixed up? Databases can include missing data elements, noisy data, inconsistent and/or multiple data elements and such. These all generate low-quality of the explored data, in general. Specifically, in this project – it affected the DM process and cause damage to activate the next DM analysis procedures well.

3. RELATED TERMS AND PREPROCESSING STAGES

Each selected data mining processing model is based on few critical data analysis aspects, which must be defined and categorized properly: data characteristics, level of detailed data and the data scope, environment of the research field, accuracy level of the data and type of the implementation to be managed. The DM world offers and equipped us with two different model types: descriptive and predictive, as mentioned (Dunham)(Han). Each approach has few leading root models and branched models for almost each root

model. So, what is the preferred model which is best-fit for our mission? Activating suitable array of decisions can eliminate immediately the improper models and ensure that we are in the "correct" and "right" environment, in our way to make the final decision for our best-fit model selection. Different kinds of DM models are proposed to us, according to these wide range of pre-checks, preprocess estimations and evaluations we need to make, to be sure that the results would be maximized: reliable and validated **scientifically**. The preprocessing stages is a strategy. It is not a recommended set of what essential or not essential to do. It is a set of MUST-BE-DONE and inevitable set of procedures, instructions, and how to-do detailed information. Just small range of degrees of freedom is permitted. Any small exception may lead to incorrectness. The article backups and presents one typical mistake, which was recognized during the research. It shows that even professional companies and development teams are not always "immune" from making such mistakes. Wrong model selection leads, of course, to total wrong results: false-distorted results, erroneous results which may cause to improper results' interpretation. In addition, confusion, distrust and mental negative aspects accompanied the entire implementation process. The line below describes the complete and proper preprocessing 4-stages, regarding the data elements. (1)Cleaning → (2)Integration → (3)Reduction → (4)Transformation & Discretization. Briefly, the *data cleaning* includes completion of missing values, removing outliers and resolve inter-consistencies. The *data integration* stage merger multiple databases and/or data cubes into one central integrated data structure. Next, the *data reduction* sub-procedure copes with data and dimensions reductions, without damaging the data properties. Finally, the data transformation & discretization procedures are required to customize the data for the next data analysis DM models, for a proper, convenient and effectively working procedures (Han et al, chapter 3). In addition, few more data mining aspects must be considered, in order to achieve a complete wide efficient way to analyze the entire data and in order to establish a correct DM model. These aspects cover the following accepted terms: Overfitting, Outliers, High Dimensionality, Integration (Dunham, p.15-16). Overfitting: Once a DM model was established, it must fit future database states. Overfitting occurs when the model does not fit future states (Dunham, p.14, p.77). Outliers: These are all the data elements which do not fit the data fields format from few possible reasons: invalid experimental error or any other validity in the measurement process. In addition, wide range of reasons can cause "noise" to the entire data set, like: range (less/greater than specific values), type (string instead of numeric values), missing values, duplicated values etc. These all cause "noise" to the entire data set and cause data distortion, which finally decrease the accuracy level of the data analysis process. High Dimensionality: conventional DB schema may be composed of many different attributes. NOT all these attributes are always needed for the main data analysis process that we want to implement. This may lead into dimensionality curse, which means that we need to specify correctly the precise attributes' set (Dunham, p. 15). Integration: Knowledge Data Discovery requests may be treated as special, unusual or one-time needs. These make them inefficient and ineffective. DM integration into traditional DBMS is certainly a desirable goal (Dunham, p.15). The data mining researchers must be aware prominently and consistently that the attribute selection can be critical to the entire analysis process. Wrong or even unintentional distinguishing between nominal and numerical attributes data objects (Han et al, p.40), from within the specific data mining models – affects immediately the results, and cause data distortion. This part is based on *attribute selection and pattern evaluation* (Han et al, p. 336, 224, 40, 264, respectively.)

4. THE FINDINGS

Can predefined array of MUST-BE-DONE actions and activities can ensure reduction and reducing the estimated error/s on one hand, and increase certainty of the results on the second hand? The answer is 'simple' BUT complicated: YES, it can be done – for sure. BUT very professional step-by-step systematic and consistent stages MUST be done. JUST very professional and experienced staff must lead and manage such projects. The staff skipped over some critical professional issues, from uncertain and uncleared reasons, which navigate into improper professional process. Part of the mistakes have been done from within the web-mining model itself, as well. In wide perspective, it seems that in spite that, the major problems were

caused by skipping over the preprocessing rather than the mistakes in the web-model itself. The list below summarizes precisely the necessary practical-implemental conclusions:

- 1) The DM preprocessing methodology requires step-by-step of proper, accurate, rechecked and monitored all the data elements in the research field. We have to "stick" to the preprocessing methodology.
- 2) The need of using normalized data is essential. Some of the models must have their input as normalized data, on the range (0,1). Otherwise – it may be ineffective to continue with the next DM-process.
- 3) Distinguishing between parameters' types is critical. Specially, the distinguishing between nominal and numerical values may affect the DM-process (Han et al, p. 40), and lead to total unexpected results. The analysis rulers (such as Gini Index, CART, Gain ratio) are based on this distinguishing (Han et al).
- 4) Cross-checking of few models' results can leverage and increase significantly the entire correctness and preciseness of the entire project. These are important as well, in the entire pre-process array decisions and considerations, and at the end, whilst all the models' work was completed, and results should be interpreted (Fauset) (Krell et al). For example, established the classification model with regression, beside neural-network model/s. This was a critical issue; a conceptual mistake, which has been done by the staff, which decided to concentrate only on one model.
- 5) From managerial point of view – Knowledge expert must follow ALL the project stages carefully, with measures steps; "heel to toe" (Chipman et al). This part was basically not implemented well during the project. Project manager with wide knowledge about DM aspects can cover and solve out this point. Regarding to this point, it is recommended to appoint a specific data expert – to the preprocessing stage itself. Thus, another cover of "two-eyes" professional programmer – would prevent such mistakes, as mentioned.
- 6) It is recommended that an expert in web-mining will be the project manager. This expert should manage the entire project work, side by side with sociologist and with information-technology. The sociologist should cover the sociology aspects, and the information-tech should cover the DM side.
- 7) Finally, the cost of the total project was significantly much higher than the expected, approximately 20% of the planned budget. The reason was the necessity to fix-up the data elements, and reactivate the DM analysis models, in order to interpret the results. "Think before you act, act proactively". Investing much THINKING at the beginning and more planning – saves a lot of money and means. Always correct, but sometimes we skip it, without meaning to.

REFERENCES

- Awad, M. and Thuraishingham, B. and Wang, L., 2009. *Design and Implementation of Data Mining Tools*, CRC Press.
- Chipman, S. and Meyrowitz, A., 1993. *Foundations of Knowledge Acquisition, cognitive Models of Complex Learning*, Kluwer Academic Publishers.
- Dunham, M.H., 2003. *Data Mining: Introductory and Advanced Topics*, Prentice Hall, Pearson Education Inc.
- Fauset, L.V., 1994. *Fundamentals of neural networks, architectures, algorithms, and applications*, Prentice Hall.
- Feelders, A. and Daniels, H. and Holsheimer, M., 2000. *Information & Management*, Methodological and practical aspects of data mining, Elsevier, Vol. 37, Issue 5, pp. 271-281.
- Han, J., and Kamber, M., 2011. *Data Mining: Concepts and Techniques*, 3rd Edition, Morgan Kaufmann.
- Hand, D. and Heikki, M. and Padhraic, S., 2001. *Principles of Data Mining*, MIT press.
- Hatzi, O., and Zorbas, N. and Nikolaidou, M. and Anagnostopoulos, D., 2014. Panhellenic Conference on Informatics, *An intelligent tool for expediting and automating data mining steps*, ACM International Conference Proceeding.
- Kaur, J. and Garg, K., 2019. Advances in Intelligent Systems and Computing, *Efficient management of web data by applying web mining pre-processing methodologies*, Vol. 731, pp. 115-122.
- Krell, M. and Wilshusen, N. and Seeland, A., and Kim, S.K., 2017. *Journal of Neural Engineering*, Classifier transfer with data selection strategies for online support vector machine classification with class imbalance, Vol.14, Issue 2.
- Lawrence, K.D. and Kudyba, S., and Klimberg, R.K., 2008. *Data Mining Methods and Applications*, CRC Press.
- Michael J.A.Berry and Gordon L., 2011. *Data Mining Techniques: For Marketing, Sales and Customer Support*, John Wiley& Sons, 1st Edition.
- Sivakumar, P., and Prakash, M., and Singaravel, G., 2015. *International Journal of Applied Engineering Research*, Efficient methods for distinction preclusion in data mining, Vol. 10, Issue 55, pp. 2212-2215.
- Uwe, L., 2012. PowerPoint presentation, *Artificial Neural Networks and Data Mining*, Hochschule Wismar University of Technology, Business and Design, (slide: Table1).
- Velasques, J.D. and Lakhmi, C.J., 2010. *Advanced Techniques in Web Intelligence – I*, Springer, SCI, Vol. 311.