https://doi.org/10.48009/1_iis_2020_147-152

Issues in Information Systems
Volume 21, Issue 1, pp. 147-152, 2020

TRUST IN LEARNING FROM BIG DATA: THE TWO SIDES OF THE SAME COIN

Dimitar Christozov, American University in Bulgaria, dgc@aubg.edu
Emanuela Mitreva, American University in Bulgaria, emitreva@aubg.edu

ABSTRACT

Building application in a way to support reaching the confidence and trust is based on the way how the user acquires, understands and interprets obtained information. In the case of Big Data information is in the form of summary statistics, aggregations, and especially important – visual forms. Producing an active knowledge, or building an informative application, requires that the user trusts the information, obtained via the Big Data processing application. The source data and the data processing application is considered often as a black box, which results in different users' attitude toward confidence in applying it. The two extreme kinds of users' acceptance and consequently attaining knowledge are: 1) Over trust: the user believes that once the information is produced by computer application it is correct, valid, and trustful, and 2) Under trust: because the user cannot observe the entire story from data generation, storing and processing, toward generating results, the information produced is doubtful and cannot automatically result in action. Both extremes are risky from point of view of the success of the application to inform the user. Optimists, over-trusting obtained information, will act according to it not critically assessing whether it is true or applicable to the problem domain. Pessimists, under trusting, may not act even if obtained information is useful, because they are unable to trace how it is created and are not ready to adopt it. The paper discusses the aspects of designing of user interface (UI) in a way to achieve confidence toward conclusions and mitigating the above risks.

Keywords: trust, machine learning, informing, users' profiling, UI architecture

INTRODUCTION

The current stage of human society is marked with terms like "information society" or "knowledge society". This describes clearly that society, in the diverse meaning of this word, is dominated by using data and information in a way to obtain knowledge. Knowledge needed to act, progress, and prosper in nowadays. Advancement of computer and communication technologies, in short IT, reached the point that every facet of human life is heavily influenced by using IT in transforming information into knowledge. This motivates, direct, and drive all kind of human activities. The last three decades, since the WWW have been introduced, changed completely the way of learning and consequently the risks of being misled changed in nature, likelihood, and impact. The magnitude and exposure of such risks are defined by the overwhelming use of IT as well as the pressure to become informed.

Three aspects need addressing: Big Data phenomenon challenged users by inability to observe data in its completeness; making decisions under the pressure of globalizing world with increasing complexity; and the growing importance of human-computer interaction to allow fast and easy capturing the essentials.

Big Data Phenomenon

Big Data era is defined by users' ability to benefit from data (see Christozov & Toleva-Stoimenova, 2015). What represents Big Data phenomenon nowadays is defined by the following aspects that influence the way human may benefit:

- Data availability: nowadays, practically, all facts associated with every event may leave a digital footprint, but not all data is well organized, reasonable structured, and easy to access, retrieve, analyzed, and present. Data is available, but accessing, exploring, and using is not always easy.
- Volume, Variety, and Velocity – the original 3Vs characteristics of Big Data: Big Data can be seen only via the use of specially designed computer application and visualization plays a critical role. Huge Volume of data requires visualizing only summary, aggregations, and statistics to represent a given data set. "Variety" refers to different formats and ways data are captured and stored and adds complexity in summarizing and presenting. "Velocity" reduces the time data is relevant and up to date. Since its original publication made

by Gartner, many authors identified other Vs as Variability, Veracity, Vulnerability, etc. which identify other sources for increasing complexity of learning from data and inspiring confidence. Any of those "Vs" challenges the use of data.
- Availability of data and data mining or machine learning application offer the opportunity to discover unknown patterns and relationships, hidden in data. But mathematics implemented in those analytical techniques use assumptions that are not always easy to justify and also to allow correct interpretation by non-professional statisticians.

Pressure and Challenges in Decision Making

Rational, data-driven decision making is becoming more complex, but decision makers have less time to comprehend options in a way to reach the needed confidence. The following lists just a few of nowadays pressures and challenges faced by managers:

- Globalization of markets – the pressure to respond to global competition require making rational, data-driven decisions. Intuition, experience or emotion may mislead decision-makers with much higher probability then it was in the past.
- Data-driven decision-making is becoming more complex because it depends on computer application mediating learning.
- Role of the trust toward the quality of obtained data, reliability of applied technology, the correctness of algorithms implemented in data analytic software, and how relevant are results to the problem is becoming critical.
- Informing, the way to build acting knowledge rely on credibility of obtained information, depends on multiple factors, divided by Z. Gackovski (Gackowski, 2006) into primary factors, as obtaining information from multiple independent sources and possibility to trace the origin of information; and secondary factors as trust toward the source, warranty, etc.

User Interface

Human-computer interaction is becoming critical. Designing user interface (UI) in a way to address specific user's attitude toward data, applications, and visualized results is becoming of high importance for the success of a Big Data application.

- Among the factors inspiring trust, Gackovski (Gackowski, 2006) recognized also the content, composition and form of presenting the information.
- The success of becoming informed depends also on information asymmetry between end-user (receiver) and the original creator of information (sender).
- User Interface (UI) is intended to provide two major services – first to ensure that the user will be able to view information by tracing in depth the sources of data, analytical algorithms and their implementation, and how results are presented; and second – to visualize results in a way suitable to given user, according to user's background, considering information asymmetry, and exploring user's attitude – whether the user is optimist, pessimist or realist.

The paper presented a concept of designing UI architecture, allowing flexibility in visualizing results of Big Data analytical application by emphasizing the user's personality via customization based on user profiling.

BACKGROUND

Trust is a key aspect of transforming information into acting knowledge. Reaching confidence in decision-making is becoming more and more critical for the business, but also in every aspect of human life today. To make a decision one needs to reach the level of confidence that this decision is correct and solves the problem. There are two general approaches in reaching confidence:

- Rational – by using known quantified measures, reflecting the state of the problem domain and build solely on available data; and
- Emotional – by using one's intuition and subjective, often qualitative, feeling regarding the correctness of decision and confidence toward invoked actions.

In practice, the two are usually combined. Even in rationally oriented decision-makers reaching 100% confidence is almost impossible to achieve and some component of emotion is needed to make an effective decision; as well, an

emotion-based decision-maker needs some pieces of evidence to justify their decision, as well as their intuition" is built on past experience.
There are at least three aspects of inspiring trust to Big Data analytics results:

- Trust to data;
- Trust to technology – algorithms and computer applications; and
- Trust to the applicability of results to the problem domain.

Data Quality

Communication, as the area of misinforming risk, is heavily influence by the form of conduct. In direct (face-to-face) communication there are many additional tools to support correct interpretation of the message as body language, environment, context, and opportunity for feedback. Indirect communication, as the Big Data case, trust cannot be generated by using additional tools, only content of the message and how the message is presented may inspire trust toward its validity. In Big Data analytics data is presented only in form of summary statistics and diagrams. According to (Gackowski, 2006), credibility of information, depends on fulfillment of several primary or secondary criteria:

- Information is based on multiple, but independent sources. There are three cases: information from all is highly correlated; complemented; or contradicting. The first two cases increase the trust, but the third decrease it.
- The reputation of providers.
- Precise, complete, original, trustful, useful, and not ambiguous information.
- Opportunity to trace the origin of the information.
- A reliable way of information capturing, storing and retrieving.
- Possibility to reproduce and replicate results.

Almost of those, inspiring trust factors are hidden in contemporary Big Data Analytic application. Emphasis is given to how to use given application by stressing on "user-friendly", "intuitive" interfaces.

Data Analytics

There are two principal categories of Data Analytics tools:

- Data Warehouses are specially designed applications to viewing data in its multidimensional nature on different levels of aggregation in generalization/specialization hierarchies.
- Data Mining tools allow discovery of "interesting" (unknown) patterns and relationships. Often algorithms implemented in Data Mining applications lay in the category of Machine Learning. Nowadays two major approaches in Machine Learning can be distinguished – Statistical Vs. so-called "Deep" learning, which explores the neural network. The two allows finding relationships within data for characterization, classification, and prediction. But while statistical learning results in a well-defined model, in principle compact and interpretable, the deep learning produces a computer tool to process data as a "black-box".

The two categories of analytical tools play different roles in decision making, while Data Warehouses are used to present data in a flexible and rich manner, Data Mining helps to find new, unknown information. From this perspective, trust toward Data Warehouse results depends solely on trust on the quality of data. Calculating summary statistics are straightforward and do not generate mistrust.

Statistical Machine Learning uses complicated techniques requiring a good understanding of their applicability on the given data set. Some of these techniques are sensitive to particular properties of data, as the independence of variables or Gaussian distribution. Nevertheless, the obtained knowledge allows competent rational decision-makers to interpret trustfully results on the problem domain.

Deep learning requires users to trust that the architecture of the neural network correctly presents the relations between characteristics of investigated objects and is complex enough to replicate reality. Looking on a neural network as a black box requires the user to believe in it by default, because it proved to be correct in numerous cases in the past. It is difficult guarantee that the obtained knowledge is useful outside of the scope of the training data set. Deep learning may serve better "emotional" decision-makers.

Results of Data Mining need further support to users to become trustful. In Statistical Learning UI may help in two ways – to show how sensitive is the given technique to a minor violation of required properties of data, and second to show how data satisfy or to what extent they violate these properties. In Deep Learning case, the UI has to be able to map the scope of training and testing data sets to data processed, as well as to provide testing success rate. In both cases, UI plays a critical role in ensuring trust in the results. Designing UI depends on the selected learning approach, the given technique, but also on the user's profile.

Data-driven decision making, in case of the Big Data, requires trusting not only data, but also trust toward the technology used to extract useful knowledge from data – learning, and trust in the interpretation of obtained results in the problem domain. Designing UI in a way to support building trust toward obtained knowledge is a challenging task. It is not solely designing "intuitive, user-friendly" interface. Such requirements address only how to explore application in an efficient way to obtained results, but do not provide support of how to use the results, and especially to support building trust that the obtained results are meaningful and useful in solving user's problem. This problem was identified by M. Buckland (see section "Variety of Task Complexity", Buckland M. 1991, page 171) thirty years ago.

In the next section, we provide a conceptual model of UI architecture intended to address these challenges.

USER INTERFACE DESIGN: A CONCEPTUAL MODEL

Based on the discussion in the previous section we can identify several components of UI architecture. The conceptual model we are proposing is based on a three-level hierarchy. The model requires generating specific user's profiles by exploring application logs and accessing other relevant logs, which captures information for all of the requests of the users, then based on that profile customize the interface. The UI, as generated by the method, shows only the information that the user is interested in in a way the user may capture it. When it comes to Big Data it is quite a challenge to visualize it, because the human perception cannot process too much data (Agrawal, Kadadi, Xiangfeng, & Adres, 2015). The model considers user's interest and competency and provide analytical results in a way to avoid overwhelming by the amount of data or by the complexity of analytical techniques. When needed a meaningful explanation is provided, as well as tool to trace origin of raw data. Figure 1 illustrates components of UI model.

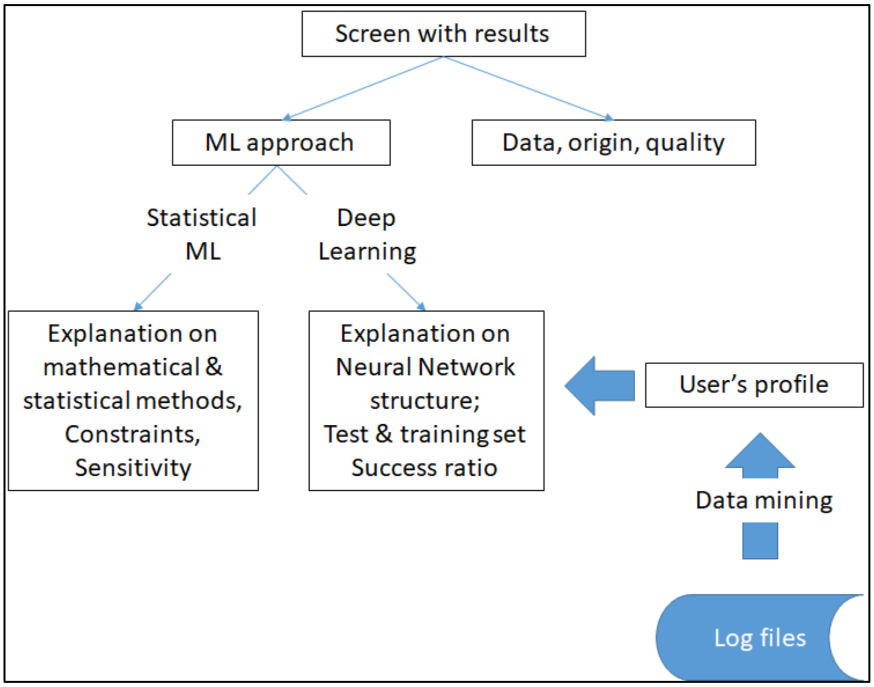
Screen with results
ML approach
Data, origin, quality
Statistical ML
Deep Learning
Explanation on mathematical & statistical methods, Constraints, Sensitivity
Explanation on Neural Network structure; Test & training set Success ratio
User's profile
Data mining
Log files

Figure 1. Customization of UI Based on User's Profiling

DISCUSSION

The architecture that we are proposing formally can be divided in:

1. Cleaning and processing logs.
2. Forming user profiles based on:
    a. Interests - a user might be interested in more detailed information about the actual algorithm that was used or just some explanation why this was the best approach that was used.
    b. Competencies – a user might be interested in a simple explanation just because they cannot understand the theory behind the actual method.
3. Defining three levels of complexity for the details for the explanation of the mathematical and statistical methods:
    a. Beginner – simple explanations on the methods used with more visualization and less mathematical details – this level can be used by either people that are not interested in the more complex details or don't have the competency to comprehend the details.
    b. Intermediate – additional details, but still with less mathematical and statistical formulas.
    c. Advanced – this level will provide the most details – visual, mathematical and statistical information.
4. The model will offer at some extend customization of the complexity of the information that is shown.

For the first two parts of the model, we could use web usage mining process – the process of extracting useful information from web pages – from logs, content or hyperlinks (Rathi & Vishwavidyalaya, 2019). The web logs can provide information about the interests of the users based on the time they spent on a particular part of the explanation of the method. But before we could be able to form the profiles of the user, the data from the web logs must be cleaned. Cleaning the data before using it will make the process of extracting user profiles more effective because around 30% - 40% of the data that is captured in the access logs is meaningful, the rest is as a result of robot requests (Marsden, n.d.) and unsuccessful requests (Manchanda, 2018). After cleaning the data, the preprocessing in the WUM includes session and user identification (Mehra & Thakur, 2018). The user behavior identification can be done using different methods like:

- The amount of time a user spent on a certain part of the explanation of the method might tell us their interests – if a user spent a little to no time on something we could either conclude they are not interested in this part or they considered it too complex for their competency. Either way, this means that the UI architecture could adapt and show only the parts that the user is interested in.
- Using path profiles to predict future user behavior (Mobasher, Dai, Luo, Sun, & Zhu, 2000).
- Clustering user sessions (Mobasher et al., 2000).

The web logs processing might give information for the user profile, but further granularity might be needed (identifications of the user sessions) to obtain the knowledge for the interests of the user.

Using just the created user profiles sometimes is not reliable enough, because the interests and the competencies of the user might change, or they would like to manual configure what information to be visible. Because of that the architecture includes also customizations using two methods:

- Offering three levels of competencies – beginner, intermediate or advanced for the user to choose from. Depending on the level different details will be shown – from just simple details for the method used if the first level is used to details about the statistics and mathematics used to mine the data.
- Offering a way to even further customization of what the UI will show – overriding the predefined three levels that were previously mentioned.

Two of the aspects in the architecture were the forming of the user profiles and creating the three levels of hierarchy, however visualization of the data is an essential tool in understanding data, because dealing with large amounts of data can lead to chaos (Syed Fiaz, Asha, Sumathi, & Syed Navaz, 2016), thus making handling big data not only an advantage, but a challenge. The proper interpreting and visualizing the data can help both a layman and a research trust in big data, but the proper visualization and tools should be used (Keahey, 2013).

Two approaches that could be used for the visualization are:

- Building samples of data (Syed Fiaz et al., 2016) with which to analyze the data.
- Creating template charts and graphs to display the data ("Shield UI," n.d.).

CONCLUSION

In this paper is offered a simple conceptual model of UI architecture, which incorporates intelligent use of accumulated in the applications' log file data. Customizing the interface based on user's competences and history of using it is expected to increase (1) user's awareness regarding application, data, and especially constraints, limitation, and applicability; and (2) to enhance the trust toward provided analytical results and confidence in making decisions based on those results. The presented conceptual model is used to design a prototype of an intelligent back-end tool for customizing user interface for complex data analytic applications. The prototype is under construction for a cloud-based e-commerce web site. It is designed to serve decision makers on tactical or strategic managerial level. The accumulated experience in applying the prototype will serve in adjusting the concept in designing customizable user interface.

REFERENCES

Agrawal, R., Kadadi, A., Xiangfeng, D., & Adres, F. (2015). Challenges and Opportunities with Big Data Visualization. *7th International Conference on Management of computational and collective IntElligence in Digital EcoSystems (MEDES)*, pp. 169–173.

Buckland M. (1991). Information and Information Systems, Praeger.

Christozov D., Toleva-Stoimenova S. (2015). Big Data Literacy - a New Dimension of Digital Divide: Barriers in learning via exploring Big Data, in *Strategic Data Based Wisdom in the Big Data Era,* editors Girard J., Berg K., Klein D., IGI Global

Gackowski, Z. (2006). Quality of Informing: Credibility Provisional Model of Functional Dependencies. *Proceedings of the 2006 InSITE Conference*. https://doi.org/10.28945/3017

Keahey, T. A. (2013). Using visualization to understand big data. *IBM Business Analytics Advanced Visualisation*, 16.

Manchanda, M. (2018). Web Usage Mining: Dynamic Methodology to Preprocessing Web Logs. *Helix*, *8*(5), 3810–3815. https://doi.org/10.29042/2018-3810-3815

Marsden, S. (n.d.). How Do Search Engine Crawlers Work? -DeepCrawl. Retrieved January 9, 2020, from *https://www.deepcrawl.com/knowledge/technical-seo-library/search-engine-crawling/*

Mehra, J., & Thakur, R. S. (2018). An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining. *International Journal of Applied Engineering Research*, *13*(2), 1227–1232. Retrieved from http://www.ripublication.com

Mobasher, B., Dai, H., Luo, T., Sun, Y., & Zhu, J. (2000). Combining Web Usage and Content Mining for More Effective Personalization. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *1875*, 165–176. https://doi.org/10.1007/3-540-44463-7_15

Rathi, P., & Vishwavidyalaya, G. K. (2019). An Efficient Algorithm for Data Pre-Processing and Personalization in Web, *International Journal of Computer Sciences and Engineering Open Access An Efficient Algorithm for Data Pre-Processing and Personalization in Web Usage Mining*. (May). https://doi.org/10.26438/ijcse/v7i5.160164

Shield UI. (n.d.). Retrieved from https://www.shieldui.com/

Syed Fiaz, A. S., Asha, N., Sumathi, D., & Syed Navaz, A. S. (2016). Visualization: Enhancing big data more adaptable and valuable. *International Journal of Applied Engineering Research*, *11*(4), 2801–2804.