International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

A Literature Review on Big Data and Time Series

Ajla Kirlic*[1], Aldin Hasovic[2]
*[1]Information Technology, American University in Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina
[2]BHANSA-BiH Air Navigation Service Agency, Sarajevo, Bosnia And Herzegovina

ABSTRACT

For people who need to make important decisions there is a lot of data that they need to handle. Huge amount of data known as big data denotes large datasets that have high velocity and variety that makes them hard to handle using some known techniques and tools. Main idea of handling big data is to provide valuable insights for decision makers to make valuable and precise decisions. Best way how to deal with big data is to use big data analytics which includes using time series methodology. This paper has a goal to go through literature that refers to big data, time series and different big data analytics methods using data mining.

Keywords : Big Data, Data Mining, Forecasting, Time Series

I. INTRODUCTION

From last decade big data is becoming one of the most interesting and important trends and has a huge potential to change the way how companies are organizing information which is useful for them, clients and potential clients to transform that data into useful business models.

If there is no way to storage data, companies would lose the ability to get valuable knowledge, information and perform different analyses to provide new advantages and opportunities. For companies, all information about customers are essential for creating new products and making their customers satisfied, so they use data as the building block to get their results.

Nowadays though advancements of internet and technology the details about customers are extent, so methods of data collection, storage capabilities and huge amounts of data become reachable. To get some useful information all data needs to be stored, and nowadays storage of data is much cheaper.

Changes in variety and size of data require the development of new methodologies for data analytics and forecasting.

On the other hand time, series forecasting model observes different values predict future values. Regression analysis often tests theories that the current data of one or more time series has the impact on the current data of another time series [1]. Time series lately is becoming very popular, a reason for that is decreasing hardware's cost and capability of processing.Time series data occurs in many areas like financial analysis [2], sensor monitoring of network [3], analysis of medical issues [4], and mining of social activity [5]. Since increasing time-stamped activity volume lately, there is the opportunity for data scientists to describe, find measure and forecast behavior of important evolutions [6].

In last few decades, communities of researchers are well-known with topic of time series, also there are some new reveals where and how time series can be used:

- ✓ If we combine big data and time series, fully automatic mining has even higher importance
- ✓ Widely usage of nonlinear modeling in different areas like economy, biology, medicine physics
- ✓ The possibility of modeling as tensors and different applications in hyperlinks, social networks, streams [7].

The idea and objective of this paper are to provide an analysis of the literature that is available on big data analytics and time series. However, there are several methods, big data tools, and technologies that can be applied and that we can discuss.

Also, literature is carefully chosen on its importance, novelty and popularity for the research and publication years are from 2000-2016 for big data and from 1940- 1990 for a time series. The reason is that big data and time series are being lately very popular topics.The data from literature review is coming from research papers published in conferences, journals and white papers mainly coming from industry corporations.

II. BIG DATA ANALYTICS

The term "Big Data" is recently been used to huge and massive data sets of structured and unstructured data that is difficult to process using old-style software techniques and databases. Large datasets of big data are organized as it shown in Diagram 1.
In most situations the size of data exceeds processing capability, but the importance of capturing, manipulating, formatting, managing big data gains companies' very important and useful insight in capturing new clients and increasing the quality of their services. On the other hand, by organizing data in better way companies are decreasing their expenses [8].

Big Data cannot be a single market; it is a management of data that evolved over the time period.

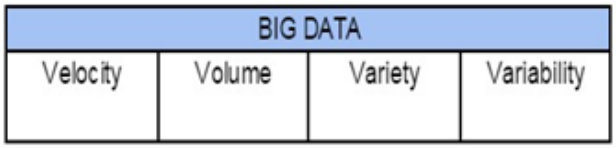

Figure 1. Properties of Big Data

Many companies are experimenting with data, but the volume of data is usually increasing going over terabytes (TB) to petabytes (PB). The idea of companies is to investigate highly detailed data in the term to discover facts and information that they didn't know about customers. As an example we can use the company which manufactures, to collect data from sensors of machines to observe detailed information and improve the process of manufacturing and avoid possible problems [9].

However, if the advanced analytic techniques are applied on big data then we have big data analytics [9]. Usually big data analytics on large sets of data reveals huge business change, but larger data sets are more difficult to manage [9].

III. TIME SERIES

Nowadays forecasting with time series are coming from the idea that if you have knowledge about the past you can predict future. Interpreting past in terms of future usually becomes main idea of time series analysis.

First thing in forecasting time series is to determine the model. The model is most often dynamic relationship among which we observe and variables that are related to our observation. Usual time series forecasting approach is the usage of regression analysis. Model in the appliance of regression includes the relationship between dependent variables and explanatory variables.

This review is interested in a modern approach to time series. First scientists in the field of linear prediction are Kolmogorov (1941) and Wiener

(1949), but their approaches were different, Kolmogorov was working in the domain of time while Wiener was working in the domain of frequency. Work of Wiener was important for the development of modern forecasting of time series; he was the first who formulated signal extraction problem. When we look back to history, systems of equations models were popular in 1950's, it was clear that models of forecasting are coming from signal extraction context [10].

However, there was the suggestion about very simple ad hoc method which was having potential to predict as well as more complicated methods [12]. Another idea was suggesting that it should be considered relevant issues more that forecasting methods and that our time would be better disbursed in considering case studies which appraise different tactics to forecasting, as contrasting comparison among methods [13].

Having this in mind, it is always easier to use modern tactics that have greater elasticity while establishing a statistical model, comparing to limiting ad hoc methods. Furthermore, an approach like this includes economic justification of forecasting method. The first example of this study was introduced by Muth in 1960 [14].

After Muth's ideas there was a new approach that the nature of the selected decomposition should be derived from an optimization model over time, and also scientist Nerlove defines UC (unobserved components) which were not straight noticeable by agents, instead agents there was a retort to the slightest mean square error [15].

In 1961 formulation of statistical Adaptive Expectations are renamed into Rational Expectations. Beside change of name new concept implied general form for the linear representation of stochastic series than the random walk plus noise model. Furthermore, this generalization allowed us more tractability to combine many assumptions according to the instructions of the economic theory [16].

Also later works highlighted the nature of the nominated disintegration should be derived from a model of optimization over time [17].
During 1980s many authors started to use unobserved components as frameworks to forecast economic series which were economic, and it was stated that the structure of time series is decomposed into trend, irregular, and seasonal components. Far ahead in 1989 same authors combined many outcomes into a book, but it is not clear their contribution besides collecting results of other researchers [18].

Harvey embraced the submission of Kalman filter for a problem of time series assessment. As an illustration, authors cast regression model with disorders in a form of state space and used Kalman filter in ducking computational worries connected with MA disturbances and overturn of their concomitant covariance matrix [19].

As an opposite of unobserved components, authors in 1981 introduced as an alternative form of decomposition. Nerlove in 1967 had a problem about suitable decomposition of time series with a signal withdrawal standpoint. It was suggested to segregate business sequences using ad hoc methods essentially and theoretically [20].

Lastly, it is quite fascinating to mention in this paper that this review highlights some of the complications of unobserved components models. Those problems are occasionally proof of identity of conditions particularly in the case in many variations [21].

IV. BIG DATA AND MAKING DECISIONS BY DATA MINING TECHNIQUES

The importance of big data when we look through decision maker's perspective is in aptitude to deliver knowledge of value and information, upon which

decision is made. The decision-making process is highly complex and has huge importance and it was researched throughout the years.

When you want to make a decision it is becoming gradually significant to have a big data. Nowadays best sources for big data are social networks, media platforms, mobile phones, different types of loyalty cards and they give excellent benefit to companies. To gain benefit it is also important to analyze data in a proper manner to get valuable awareness and the best way is to observe results from real time and affluence of historic data generated through customer performances, processes of production, etc [22].

Furthermore, most of the organizations are well organized to analyze data which is internal through inventory, sales, and shipments, but they need to analyze external data like general market, customer market and chains of supply, use and knowledge about big data provide satisfactory results. Since types of unstructured data are rapidly increasing, it becomes essential to make more knowledgeable decisions based on significant implications from the data [23].

Some authors believe that Data Mining techniques can give a hand in forecasting with Big Data [24], also it is important to mention that Data Mining in past was usually used on data which is static as divergent to time series [25].

During the recent financial crisis there are some beliefs that financial models for Big Data which were accepted were incapable dealing with huge quantities of data that was inputted into systems, and for that reason, results were giving imprecise forecasts [26].

There are varied gains through forecasting using Big Data. Nowadays there is augmented research into consuming Big Data for obtaining precise weather, in fact, weather forecasting has developed as one of the main beneficiaries of Big Data, but forecasts are still imprecise outside a week [27].

We will mention some related fields and topics where the Big Data and Data mining techniques are giving great results. Examples of those fields are Finance and Economics, dynamics of population, energy and environment, crime, media, medicine etc.

Scientists in economics are exploring Big Data for predicting variables of economic. It is used Dynamic Factor Model (DFM) by Camacho and Sancho [28] which is based on methodology presented to predict big data including dispersal indexes. Frequently used models for forecasting with big data are factor models introduced by Stock and Watson [29] and DFM models are just extension of factor models.

Finance use big data forecasting with transaction values of stock merchandized on the London Stock Exchange after cleaning data for outliers [30].

Crime is interesting area for data mining, good example is detection of fraudulent behavior in China telecom where K-means and two step clustering algorithms are used [31].

It is showed that using huge amount of youtube videos in Singular Value decomposition can be used for forecasting video patterns. Researchers found that Hierarchical Clustering can provide very efficient forecasting in media [32].

Forecasting with big data in environment is mostly referring to weather forecasting. Sigrist et al. exploits Stochastic Advection Diffusion Differential Equation to expand prediction for north Switzerland using big data. It is proved that usage of this model gives better results than raw forecasting with numerical model [33].

Support vector Machines together with neural networks are used for forecasting big data for electricity consumption in China [34]. Wavelet transform (WT) usage together with linear regression, GARCH and Multilayer Perceptron

(MLP) helped to forecast gas price and electricity demand of UK [35]. On the other hand it is appraised the usage of Exponential Smoothing and ARIMA models in mixture with a model formation consultant to forecast energy request using Big Data from an energy area [36].

V. CONCLUSION

Lately Big Data is representing very innovative topic for researchers but also for the companies who wants to develop. In this review, we have focused on forecasting with Big data, to extraordinary opportunities and benefits.

Today we live in era of information; data is being produced with high velocity and volume, and within them we have some valuable knowledge and information that needs to be utilized and extracted. Analytics of big data can be applied to important changes in business and decision making, by applying progressive techniques on big data.

Benefits of handling big data can help in many sectors like manufacturing, health care, retail etc…We hope that this paper will give an idea to others how to handle big data with time series. Any technology and big data method if it is applied in a proper way can give many innovations and benefits to decision makers but also to customers. Nevertheless dealing with big data and time series is extremely hard because big data requires propped handling, integration, storage, processing, analyzing etc. Our future researches can focus on giving roadmap to management of big data.

Finally, we believe that big data and time series have high significance on data overflow and can gain many benefits in many areas, on technological, as well as on humanitarian levels and scientific.

VI. REFERENCES

[1]. Imdadullah. "Time Series Analysis". Basic Statistics and Data Analysis. itfeature.com Retrieved January 2014

[2]. Y. Zhu and D. Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In VLDB, 2002.

[3]. S. Papadimitriou and P. S. Yu. Optimal multi-scale patterns in time series streams. In SIGMOD, , 2006.

[4]. E. J. Keogh, S. Chu, D. Hart, and M. J. Pazzani. An online algorithm for segmenting time series. In ICDM, 2001.

[5]. M. Mathioudakis, N. Koudas, and P. Marbach. Early online identification of attention gathering items in social media. In WSDM , 2010.

[6]. J. Leskovec, L. Backstrom, and J. M. Kleinberg. Meme-tracking and the dynamics of the news cycle. In KDD , 2009.

[7]. Y. Sakurai, Y. Matsubara, C. Faloutsos. Mining and Forecasting of Big Time-series Data

[8]. Kubick, W.R.: Big Data, Information and Meaning. In: Clinical Trial Insights, (2012)

[9]. Russom, P.: Big Data Analytics. In: TDWI Best Practices Report, (2011)

[10]. A. Kolmogorov (1941) "Interpolation and extrapolation von stationaren Zufalligen Folgen," Bulletin of the Academy of Sciences (Nauk), USSR, Ser. Math

[11]. N. Wiener (1949) The extrapolation, interpolation and smoothing of stationary time series with engineering applications, Wiley: New York

[12]. S. Makridakis, M. Hibon (1979) "Accuracy of forecasting: an empirical investigation (with discussion)," Journal of the Royal Statistical Society A

[13]. P. Newbold (1983) "The competition to end all competitions," Journal of Forecasting, 2(3),

[14]. Muth (1960) "Optimal properties of exponential weighted moving average forecasts," Journal of the American Statistical Association

[15]. M. Nerlove (1967) "Distributed lags and Unobserved Components in economic time series," Ch.6 in Ten Economic Studies in the Tradition of Irving Fisher, W. Fellner et. al. eds., New York: John Wiley & Sons.

[16]. Muth (1961) “Rational expectations and the theory of price movements,” Econometrica,
[17]. Nerlove and Grether (1970) “Some properties of “Optimal” seasonal adjustment,” Econometrica
[18]. Harvey (1989) Forecasting, structural time series models, and the Kalman filter, Cambridge University Press.
[19]. A.C. Harvey, G. Gardner, G. Phillips (1980) “An algorithm for exact maximum likelihood estimation by means of Kalman filtering,” Applied Statistics, 29
[20]. M. Nerlove (1967) “Distributed lags and Unobserved Components in economic time series,” Ch.6 in Ten Economic Studies in the Tradition of Irving Fisher, W. Fellner et. al. eds., New York: John Wiley & Sons.
[21]. J. Ledolter (1984) “Comments on ‘A unified view of statistical forecasting procedures’ by A.C. Harvey,” Journal of Forecasting
[22]. Cebr: Data equity, Unlocking the value of big data. in: SAS Reports(2012)
[23]. Economist Intelligence Unit: The Deciding Factor: Big Data & Decision Making. In: Capgemini Reports,(2012)
[24]. Rey, T., and Wells, C. (2013) Integrating Data Mining and Forecasting. OR/MS Today, 39(6).
[25]. Berry, M. (2000) Data Mining Techniques and Algorithms. John Wiley and Sons. 14 Biau, O., and D’Elia, A. (2009). Euro Area GDP Forecasting using Large Survey Datasets. A random forest approach.
[26]. Cukier, K. (2010). Data, data everywhere. The Economist.
[27]. Silver, N. (2012). The Signal and the Noise: The Art and Science of Prediction. Penguin Books, Australia.
[28]. Camacho, M., and Sancho, I. (2003). Spanish Diffusion Indexes. Spanish Economic Review
[29]. Stock, J. H., and Watson, M. W. (2006). Forecasting with many predictors. In Handbook of Economic Forecasting, Elliott, G., Granger, C. W. J., Timmermann, A. (eds). Elsevier: Amsterdam
[30]. Alessi, L., Barigozzi, M., and Capasso, M. (2009). Forecasting Large Datasets with Conditionally Heteroskedastic Dynamic Common Factors. Working Paper No. 1115, European Central Bank.
[31]. Wu, S., Kang, N., and Yang, L. (2007). Fraudulent Behaviour Forecast in Telecom Industry Based on Data Mining Technology. Communications of the IIM
[32]. Gursun, G., Crovella, M., and Matta, I (2011). Describing and Forecasting Video Access Patterns. In: INFOCOM ’11: Proceedings of the 30th IEEE International Conference on Computer Communications, IEEE, 2011
[33]. Sigrist, F., Kunsch, H. R., and Stahel, W. A. (2012). SPDE based modeling of large space-time data set
[34]. Wang, X. (2013). Electricity Consumption Forecasting in the Age of Big Data. Telkomnika
[35]. Nguyen, H. T., and Nabney, I. T. (2010). Short-term Electricity Demand and Gas Price Forecasts using Wavelet Transforms and Adaptive Models
[36]. Fischer, U., Schildt, C., Hartmann, C., and Lehner, W. (2013). Forecasting the Data Cube: A Model Configuration Advisor for Multi-Dimensional Data Sets. In: IEEE 29th International Conference on Data Engineering (ICDE)