BVG Trust

J. Sci. Trans. Environ. Technov. 2018, 11(3) : 155-160

Scientific Transactions in Environment and Technovation

Big Data and Topic Models – A Definitive study for emerging research

https://doi.org/10.56343/STET.116.011.003.008
http://stetjournals.com

R.Manjupargavi*, R.Bhuvanapriya and R.Vijayalakshmi

*Department of computer science, S.T.E.T Women's college,Sundarakottai,mannargudi,Tamil Nadu.
Department of computer science, S.T.E.T Women's college, Sundarakottai,mannargudi,Tamil Nadu.
Department of computer science, S.T.E.T Women's college, Sundarakottai,mannargudi,Tamil Nadu.

Abstract

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying and information privacy. Big Data is one of the latest emerging topics in the field of business information systems and is marketed as being the key to companies future success. Many analytic solutions are offered by IT companies to help other businesses with the flood of data that is generated within and outside of a company. Despite the extensive use of the notion of Big Data for marketing purposes, there is no common understanding of how to characterize the elements of the Big Data concept. The authors contribute to the clarification of this concept with a methodologically enriched literature review by deriving characteristic dimensions from existing definitions of Big Data.

Key words: Big Data; literature review; topic models.

Received : July 2017 Revised and Accepted : January 2018

INTRODUCTION

Big data means really a big data, it is a collection of large datasets that cannot be processed using traditional computing techniques. Big data is not merely a data, rather it has become a complete subject, which involves various tools, techniques, and frameworks. In this era, data are continuously acquired for a variety of purposes. Hadoop is an Apache open source framework written in Java that allows distributed processing of large datasets across clusters of computers using simple programming models. Big Data is data whose scale, diversity and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it. Big data is a voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Although big data doesn't refer to any specific quantity the term is often used when speaking about petabytes and exabytes of data etc. But today Big data can describe with 3Vs: the extreme Volume of data, the wide Variety of types of data and the Velocity at which data is traversing. As Big data takes too much time and costs too much money to load into a traditional relational database for analysis. So, new approaches to storing and analyzing data have emerged which rely less on data schema and data quality. Since 2000, data generation has been growing rapidly from various sources, such as Internet usage, mobile devices and industrial sensors in manufacturing (Hilbert and Lopez, 2011). As of 2011, these sources were responsible for a 1.4-fold annual data growth (Manyika *et al*., 2011). Furthermore, the storage and processing of the data have become less expensive and easier due to technological developments, such as distributed and in-memory databases that run on commodity hardware and decreasing hardware prices (Armbrust *et al*., 2010). The resulting massive influx of data has inspired various notions about the future of information science, with the most popular notion being Big Data. In practice, the notion of Big Data is widely used and marketed as the key for future companies' success; thus, IT companies have built many products to capitalize on this concept. The fields of application are diverse, and they include but are not limited to, location studies (IBM,2011), dynamic pricing (Dignan, 2012) and scenario development (Lavalle *et al*., 2011). The proposed areas of application incorporate most industries and corporate functions. Both governments and the health-care sector have also been paying attention to the development (Wesis and Zgorski, 2012; Feldman e*t al*., 2012). Moreover, the recent developments in this field pose a challenge regarding the demand for data scientists. According to the results of a study by McKinsey Global Institute, which focuses on the US labor market, one challenge of the Big Data concept will be the shortage of data scientists, which

*Corresponding Author :
email: mprithishni5@gmail.com

P - ISSN 0973 - 9157
E - ISSN 2393 - 9249

has been projected to be between 140,000 and 190,000 by 2018 (Manyika *et al*., 2011).

Existing Definitions and Describing Dimensions

The authors have found a continuous increase in the number of publications that address Big Data in scientific databases, such as Scopus, every year since the beginning of the 21st century. This increase culminated in a sharp increase in publications in 2010 for these databases. This development can also be found when analyzing the interest in the keywords "Big Data" with Google Trends. During the literature review, the authors noted that similar to other emerging topics, no common understanding of the notion of Big Data exists (Madden, 2012). Existing meta-studies of the Big Data research field have divided the Big Data developments into Business Intelligence and Analytics 1.0 - 3.0, developing different maturity levels and describing key characteristics for each level (Chen *et al*., 2012). Pospiech & Felden review the current literature on Big Data, revealing a focus on the technical perspective of data provisioning (Pospiech and Feldman *et al*., 2012). With regard to the different ranges of use, the phrase Big Data entails a potential misunderstanding because it is used both to describe the size of the processed datasets and to describe the entire Big Data concept. The authors suspect that the sheer breadth and depth of the topic hinders the formulation of a consistent definition. The number of definitions for Big Data increases with the increasing number of publications in the field. Therefore, the authors first searched for definitions of Big Data in existing Big Data-related review publications that originated from top-ranked journals and conference proceedings. With regard to the number of definitions, the list (Table 1) is not exhaustive, but the definitions included in this table cover the most important aspects and illustrate the different dimensions of Big Data. The definition by (Cuzzocrea *et al*., 2011) is aimed at the characteristics of the generated data, containing both the amount and structure of the data, complemented with naming exemplary data sources. (Bizer *et al*., 2011) enrich the data characteristics by additional attributes, such as the scope, target, and structure of the data, addressing data heterogeneity in a "Big Data world". With regard to the data characteristics, (Jacobs, 2009) focuses solely on the amount of data and adds the aspect of the method. (Chen *et al*., 2012) include the aspect of the method in terms of analysis as well and add IT infrastructure topics, such as storage and processing purposes. Furthermore, their definition enhances the dimension data characteristics by naming a selection of data sources. The definition by (Madden, 2012) incorporates both data characteristics and infrastructure (tools), which is extended by (Manyika *et al*., 2011) with the aspect of the method. Both definitions, along with that of (Jacobs, 2009), emphasize the excessive demand of the current IT infrastructure to handle the changes in the data characteristics. This description is in contrast to one of the early definitions (Diebold, 2003), who states that the availability of the enormous amount of data is a result of the "advantages in recording and storage technology", which suggests a change in the requirements regarding the IT infrastructure. In summary, three views on Big Data can be derived from the presented definitions. The named aspects of data characteristics (amount and structure) and sources can be merged into a data dimension. The named tools and databases that are required to store and manage data can be combined to an IT infrastructure dimension. The data processing for analysis purposes can be embraced into a methods dimension. The latter two dimensions are similar to the analysis by (Pospiech and Feldmen, 2012). The results of this deductive approach are used as a basis for the further exploration of the Big Data concept. By analyzing the abstracts of highly ranked articles and conference proceedings that are associated with Big Data, the derived dimensions are validated in the first step and then enriched with contained topics.

Validation using topic models

We derived the relevant dimensions of Big Data in terms of the data, IT infrastructure, and methods for analysis purposes by using a deductive approach that analyzes existing definitions. To validate these dimensions, a structured literature review following (Webster and Watson, 2002) is applied. We enrich this approach by a methodological component from the field of text mining. The identified publications are processed in a two-step approach. First, we validate the derived dimensions by applying topic models on all of the identified publications, and then, we enrich the individual dimensions by applying topic models on dimension-specific publications. Topic models are hierarchical probabilistic models that have their origin in the field of machine learning. Topic models have been broadly applied, especially in the field of literary analysis (Titov and McDonald, 2008). The basis for topic models is unstructured data, which can be text, pictures, or videos (Wang and Mori *et al*, 2009). In this paper, we focus on the abstracts of publications related to Big Data. The individual documents are merged into a corpus that is the input for the analysis. Following the assumptions of topic models, a topic is defined by the appearance of certain words; therefore, topics can be represented as probability distributions over words. Furthermore, a document is viewed as a mixture of topics; thus, the topics within a document can be

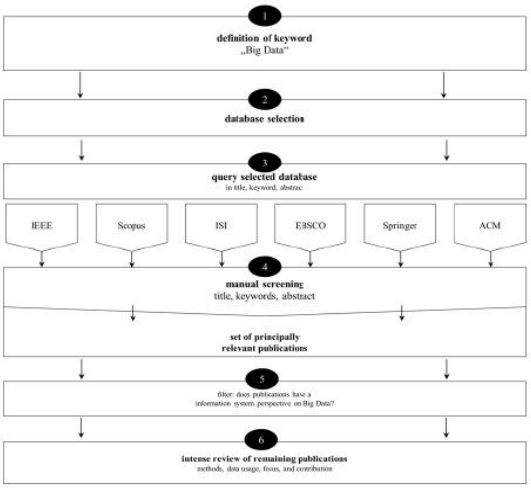
represented with a probability distribution over the topics. Following these assumptions, a document can be generated by using the distribution over the topics and, depending on the chosen topic, a distribution over the words to select the words for the document. Applying topic models on a corpus, this generative process is reversed to estimate both of the distributions with the help of a machine learning technique. Among the different estimation approaches, Latent Dirich *All*ocation (LDA) (Blei *et al*., 2003) has been widely applied for similar purposes and enhanced as Correlated Topic models (Blei and Lafferty, 2007) or Dynamic Topic Models (Blei and Lafferty, 2006). LDA, which is applied for the analysis of the publications on Big Data, is implemented as follows: The probability of an occurrence of a word w depends on the respective topic T in a document with zipping as a latent variable that describes the original topic of the ith word. P(wi)=ÓT j=1P(wi | zi = j)P(zj = j) Therefore, P(w | z) represents the relevance of a word for a topic, and P(z) is the occurrence of the topic within the document. Thus, to identify the topics, the a-priori distributions over topics and words per topic are defined followed by the fitting of the distributions based on the analyzed corpus with the help of LDA. Different algorithms have been applied to apply the model estimation process, including Variational Bayesian Inference (Blei *et al*.,2003) and the Markov Chain Monte Carlo based method Collapsed Gibbs sampling (Griffiths and Steyvers, 2004), which has been applied during the described analysis.

Data selection and preprocessing

For the analysis, we searched for papers from scientific journals and conference proceedings from the research field of computer science since the year 2010 that contained the notion “Big Data” in the title, abstract, or keywords, which led to an initial database of 1,322 publications. The resulting publications have been scanned manually, and papers were removed if they i) belong to conference workshops, except when they contribute considerably to Big Data research, ii) are keynote-related paper editorials, or iii) had content that did not belong to the field of Big Data. Furthermore, duplicates were removed. This selection process finally resulted in a database comprising 248 documents. The abstracts of these papers have been preprocessed in terms of removing stop words and the words “big” and “data” to allow the subject of the study to not become part of the analyzed corpus and therefore falsify the results due to the word frequencies in the abstracts. Furthermore, word stemming has been executed via Porters stemming algorithm (Porter, 1980). The resulting text has been analyzed using topic models, for which the results will be explained in the next section.

Analysis of the overall database using topic models

The dimensions derived from the existing definitions of Big Data can partly be found and enriched within scientific publications on Big Data (mp = 0.76). As a first step, the words of each resulting topic were analyzed regarding i) how far they can be assigned to the derived dimensions and ii) whether they account for additional subjects compared with only Big Data to determine whether these words resulted from a dispersion of a certain topic throughout the computer science discipline. If a field is marked with N / A, the probability of the word occurring did not differ significantly from the following words, meaning that they are not representative of this topic. Therefore, these words were not considered further. The displayed number of topics has been determined according to the Harmonic mean method (Griffiths and Steyvers ,2004) and in consideration of the low number of analyzed abstracts. The words of topic 1 can be assigned to the dimension IT infrastructure. Specifically, the words MapReduce and Hadoop account directly for the aspect of technologies within the concept of Big Data. The programming framework MapReduce, which was developed by Google, and its open source implementation, Hadoop, have been designed to process voluminous data. Both aspects contributed to the rise of distributed, scalable systems within the development of Big Data applications (Dean and Ghemawat, 2008). The relevance of MapReduce and Hadoop explains the word parallel as analysis tasks that can be computed in parallel. The following figure illustrates the Literature Research Process. The appearance of performance results of publications, that focus on the performance improvement of a Hadoop cluster for certain analysis purposes (Gu and Gao, 2012) or general performance improvements based on data locality Hammoud and Sakr, 2011). The aspect of efficiency correlates with performance, and queries correlate with databases in general.



Although the last four words cannot be assigned exclusively to Big Data, they are relevant in the MapReduce / Hadoop context. Similar accounts for the word cloud, which has gained in relevance in general within the computer science discipline, nonetheless has contributed to the development of Big Data as a scalable storage environment as well (Abbadi *et al*., 2011; Ari *et al*., 2012). The following table illustrates the result of the topic models of the overall corpus.

Topic 1	Topic 2	Topic 3
mapreduce	algorithm	research
performance	method	search
processing	graph	information
system	experiment	analysis
computing	problem	social
parallel	accuracy	computing
efficiency	parameter	N/A
cloud	approximate	N/A
queries	N/A	N/A
hadoop	N/A	N/A

The words of the third topic do not fit completely with the remaining data dimension; they have a generic character of possible applications in the field of data analysis. Research results from the increasing relevance of Big Data in natural science (Karlsson *et al*., 2012).The combination of search, information, analysis, social, and computing indicates the computation based analysis of a social environment; however, the generic character of the contained words will be analyzed in the next section in more detail. These results can be validated by analyzing a sample set from the included databases, containing 105 publications both from proceedings and journals of the field computer science which is presented in the following table that illustrates the validation topics.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
programs	process	algorithms	performance	data
code	management	queries	memories	mined
analysis	development	results	system	behavior
semantics	creation	complexity	caches	image
programming	governance	set	disk	digital
language	services	class	control	research
language	studies	N/A	monitoring	application
verification	knowledge	N/A	drivers	analysis
compiler	shared	N/A	power	techniques

Analysis on the dimensional level

The identified publications have been screened manually as a first step and were assigned to the data, IT infrastructure, and methods dimensions by two scientists according to the approach by (Salipante *et al.,* 1982). The data dimension contains publications that target both the characteristics and sources of the analyzed data itself as well as privacy issues. The IT infrastructure dimension inherits papers that focus primarily on the software and hardware for the processing of large amounts of data, which correspond to the topic model results in the previous section. The methodology dimension contains papers that focus on the testing and development of methods for the analysis of Big Data. After a first run, it became clear that i) the derived dimensions do not cover the recent publications entirely. Therefore, the application dimension has been added, which inherits papers that focus on the application of recent Big Data techniques in a business context. The authors chose the notion application because it appears in the definitions of two analyzed definitions (Chen *et al*., 2012; Cuzzocrea *et al*., 2011). Furthermore, ii) it became clear that a clear delineation is not always possible; therefore, publications have been assigned to more than one dimension, if needed. The results of the assignments can be found in Table III. Following the results, recent publications have focused on the infrastructure aspect, followed by methods and applications. There were a total of 12 papers that targeted data-relevant topics, illustrating this topic did not come up as a separate topic. Again, the results are tested with respect to the extent to which they account solely for the Big Data

Topic 1	Topic 2	Topic 3
cloud	queries	network
computing	database	social
cluster	stores	results
mapreduce	search	latency
processing	analysis	traffic
parallel	research	N/A
hadoop	index	N/A
distributed	processing	N/A
platform	prototype	N/A
N/A	framework	N/A

concept. The following table gives a detail list of topics in it infrastructure dimension.

Classification of the Results

By using topic models to analyze the publications on a dimensional level, the dimensions could be enriched by the identification of subtopics. To reveal blind spots, the topics identified within the analyzed publications are now transferred from the dimensions derived into the corresponding process steps within a fitted generic data analysis process (Salipante *et al*., 1982) as shown in Figure 2 and enriched with related publications from

the analyzed corpus. Due to space considerations, a selection of the analyzed publications is presented. Compared to the assignments of publications to the dimensions, several publications do not consider only one step. Step 1, data selection, is not covered in publications within the analyzed corpus, which is surprising because the selection of adequate data sources has a major influence on the latter analysis results. The data dimension in step 2 focuses on the aspect of data consistency when processing data from different sources (chute, 2012) and algorithm-based data privacy protection, which is of special interest in the field of patient data. The IT infrastructure within the step 2 inherits, amongst the different Hadoop implementations, the aspect of data import into a Hadoop cluster. The methodological focus in step 2 is reduced to the method detection of duplicates within databases. Step 3, data analysis, accumulates most of the publications. Data security-oriented analysis frameworks can be found within the data dimension. The IT infrastructure dimension of step 3 is dominated by Hadoop advancements that focus, among other topics, on performance improvements. According to the many Hadoop publications in the field of IT infrastructure, the methods dimension is dominated by parallel computing, which is primarily based on the MapReduce framework, targeting such aspects as load balancing or energy consumption optimization. In addition to frameworks for parallel computation, the development of analytical frameworks for certain types of data are within the focus in terms of i) the network analysis, applying graph theory approaches to analyze social networks or ii) real-time analysis for streaming data (Ari *et al*., 2012). The first application-oriented publications can be found within the data analysis step; these publications contain algorithms for intrusion detection, recommender systems, and social media analysis. Publications with result interpretation can be found in the application dimension, targeting the application of usage data from web search and ontology-based data access approaches in clinical environments.

CONCLUSIONS

In this paper, we derived dimensions for Big Data based on existing Big Data-related definitions. The resulting dimensions are data, IT infrastructure, and methods of data analysis. These results could be partially supported and enriched by analyzing the abstracts of papers that refer to Big Data with topic models, which has proved to be a meaningful approach to identifying and enriching topics within this research field. By reviewing the literature and describing recent developments within the derived dimension, it becomes apparent that i) distributed, parallel computing and the Hadoop ecosystem in particular play a major role in IT infrastructure-related research. Hadoop as the open-source implementation of the MapReduce framework for distributed computation can be found within the methods dimension; another topic inherits (social) networks analysis using graph theory as a result of the increasing importance of social networks for research purposes. The topics that result from applications-related publications are not sufficiently distinct and thus do not contribute to a better understanding of the Big Data concept. Furthermore, ii) the analysis of the publications reveals that the data dimension, which could be found in the existing definitions, is not yet a relevant part of the existing publications in the Big Data field. The existing publications address privacy or data security issues. After transferring the results to a generic data analysis process, it becomes apparent that both process steps in the forefront of data analysis (data selection) and subsequent to the analysis (result visualization / interpretation and derivation of actions) are not yet covered by recent publications, even though they have been identified as relevant aspects (Chen *et al*., 2012 ; LaValle *et al*., 2011). Consequently, in addition to further research in the dimensions of IT infrastructure, for example, further developments in Hadoop implementations and alternative systems, research can be performed in the data dimension, that multiple research topics that concern the data selection and gathering process from different sources or the identification of valuable data within the gathered datasets. The target group that is specific to the visualization of the analysis results as well as the integration of the decision-making process will be relevant to future research topics.

REFERENCES

Abbadi, S., Amr El. Argawal, Divyakant and Das, 2011. Big Data and Cloud Computing : Current State and Future Opportunities, in Proceedings of the International Conference on Extending Database Technology, New York, New York, USA,P. 530–533.

Ari, I., Olmezogullari E. and Celebi O. 2012. Data stream analytics and mining in the cloud, in IEEE 4th International Conference on Cloud Computing Technology and Science, P. 857–862. https://doi.org/10.1109/CloudCom.2012.6427563

Armbrust, M., Stoica, I., Zaharia, M., Fox, A., Griffith, Joseph, R.A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D. and Rabkin, A. 2010.A view of cloud computing. Communications of the ACM, vol. 53, no. 4, P.50–58. https://doi.org/10.1145/1721654.1721672

Bizer, C., Boncz, P., Brodie, M.L. and Erling, O. 2011. The meaninful Use of Big Data: Four Perspectives. SIGMOD, vol. 40, no. 4, P.56–60. https://doi.org/10.1145/2094114.2094129

Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent Dirichl*et a*llocation.The Journal of Machine Learning Research,

vol. 3, P. 993–1022. [Online]. Available: http://dl.acm.org/citation.cfm?id=944937

Blei, D.M. and Lafferty, J.D. 2007. A correlated topic model of Science.Annals of Applied Statistics, vol. 1, no. 1, P. 17– 35. https://doi.org/10.1214/07-AOAS114

Blei, D.M. and Lafferty, J. 2006. Dynamic topic models, in Proceedings of the 23rd international conference on Machine learning - ICML '06. New York, New York, USA: ACM Press, 2006, P. 113–120. https://doi.org/10.1145/1143844.1143859

Chen, H., Chiang, R.H.L. and Storey, V.C. 2012. Business intelligence and analytics: From big data to big impact," MIS Quarterly, vol. 36, no. 4,P. 1–24. https://doi.org/10.2307/41703503

Chute, C. 2012. Obstacles and options for big-data applications in biomedicine: The role of standards and normalizations, in International Conference on Bioinformatics and Biomedicine. PMid:22779038 PMCid:PMC3392064 https://doi.org/10.1109/BIBM.2012.6392651

Cuzzocrea, A., Song, I. and Davis, K. 2011. Analytics over LargeScale Multidimensional Data: The Big Data Revolution!" in DOLAP '11 Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP. New York, New York, USA: ACM,P.101–103. (Online) Available: http://dl.acm.org/citation.cfm?id=2064695 https://doi.org/10.1145/2064676.2064695

Dean, J. and Ghemawat, S. 2008. MapReduce : Simplified Data Processing on Large Clusters Communications of the ACM, vol. 51, no. 1, P. 107–113. https://doi.org/10.1145/1327452.1327492

Diebold, F.X. 2003. Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting," in Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume III, P.115–122.

Dignan, L. 2012. Sears eyes big data for dynamic pricing, cost savings.(Online). Available: http://www.zdnet.com/sears-eyes-big-data-for-dynamic-pricing-cost-savings-7000004058/

Feldman, B., Martin, E.M. and Skotnes, T. 2012. Big Data in Healthcare Hype and Hope.Dr. Bonnie 360, Tech. Rep.

Gu, C. and Gao, Y. 2012.A Content-Based Image Retrieval System Based on Hadoop and Lucene,in Second International Conference on Cloud and Green Computing, P.684– 687. https://doi.org/10.1109/CGC.2012.33

Griffiths, T.L. and Steyvers, M. 2004. Finding scientific topics.Proceedings of the National Academy of Sciences of the United States of America, vol. 101, P. 5228–5235. PMid:14872004 PMCid:PMC387300 https://doi.org/10.1073/pnas.0307752101

Hammoud, M. and Sakr, M.F. 2011. Locality-Aware Reduce Task Scheduling for MapReduce,in 2011 IEEE Third International Conference on Cloud Computing Technology and Science. IEEE, Nov, P. 570–576. https://doi.org/10.1109/CloudCom.2011.87

Hilbert, M. and Lopez, P. 2011. The world's technological capacity ´ to store, communicate and computer information Science (New York, N.Y.), vol. 332, no. 60, P. 60–65. PMid:21310967 https://doi.org/10.1126/science.1200970

IBM, 2011. Vestas - Turning climate into the capital with Big data. IBM Corporation, Armonk, NY, Tech. Rep.

Manyika, J., Chui, Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, Tech. (Online). Available: http://www.mckinsey.com/Insights/MGI/Research/Technologyand Innovation/Big data/The next frontier for innovation.

Jacobs, A. 2009. The pathologies of big data, Communications of the ACM, vol. 52, no. 8, P. 36. (Online). Available: http://portal.acm.org/citation.cfm?doid=1536616.1536632 https://doi.org/10.1145/1536616.1536632

Karlsson, J., Torreno, O., Ramet, D., Klambauer, G., Cano, M. and Trelles, O. 2012. Enabling Large-Scale Bioinformatics Data Analysis with Cloud Computing, in 10th International Symposium on Parallel and Distributed Processing with Applications (ISPA), 2012, P. 640–645. https://doi.org/10.1109/ISPA.2012.95

LaValle, S., Lesser, E. and Shockley, R. 2011. Big Data, Analytics and the Path From Insights to Value.MIT Sloan Management Review, vol. 52, no. 2, P.21–31.

Madden, S. 2012. From Databases to Big Data," IEEE Computing, vol. 16, no. 3, P. 4–6. https://doi.org/10.1109/MIC.2012.50

Pospiech, M. and Felden, C. 2012. Big Data A State-of-the-Art," in Americas Conference on Information Systems AMCIS. P.22.

Porter P.F. 1980. An algorithm for suffix stripping,Program, vol. 14, no. 3, P. 130–137. https://doi.org/10.1108/eb046814

Salipante, P., Notz, W. and Bigelow, J. 1982. A matrix approach to literature reviews, Research in organizational behavior, vol. 4, P. 321–348.

Titov, I. and Mc Donald, R. 2008. Modeling online reviews with multi-grain topic models, in Proceeding of the 17th international conference on World Wide Web WWW 08, P. 1–15. https://doi.org/10.1145/1367497.1367513

Wang, Y. and Mori, G. 2009. Human action recognition by semilatent topic models,.IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 10, P. 1762–1774. PMid:19696448 https://doi.org/10.1109/TPAMI.2009.43

Weiss, R. and Zgorski, L.J. 2012. Obama Administration Unveils Big Data Initiative: Announces $200 Million in New R&D Investments.

Webster, J. and Watson, R.T. 2002. Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly, vol. 26, no. 2, P.13–23.