Big-Learn: Towards a Tool Based on Big Data to Improve Research in an E-Learning Environment

Karim Aoulad Abdelouarit
Laboratory of LIROSA
Faculty of Sciences
Tétouan, Morocco

Boubker Sbihi
Laboratory of LIROSA
Ecole des Sciences de l'Information
Tétouan, Morocco

Noura Aknin
Laboratory of LIROSA
Faculty of Sciences
Tétouan, Morocco

***Abstract*—In the area of data management for information system and especially at the level of e-learning platforms, the Big Data phenomenon makes the data difficult to deal with standard database or information management tools. Indeed, for educational purposes and especially in a distance training or online research, the learner that uses the e-learning platform is left with a heterogeneous set of data such as files of all kinds, curves, course materials, quizzes, etc. This requires a specialized fusion system to combine the variety of data and improve the performance, robustness, flexibility, consistency and scalability, so that they can provide the best result to the learner The user of the e-learning platform. In this context, it is proposed to develop a tool called "Big-Learn" based on a technique to integrate the mixing of structured and unstructured data in one data layer, and, in order to facilitate access more optimal search relevance with adequate and consistent results according to the expectations of the learner. The methodology adopted will consist initially in a quantitative and qualitative study of the variety of data and their typology, followed by a detailed analysis of the structure and harmonization of the data to finally find a fictional model for their treatment. This conceptual work will be crowned with a working prototype as a tool achieved with UML and Java technology.**

***Keywords*—big data; e-learning; data structuring; learning; digital pedagogy**

I. Introduction

The users of the Internet, whether it was humans, programs or services provide every day enormous amount of data that become so difficult to manage and deal with traditional database management tools. The massive exploitation of such information data herein is called Big Data or massive volumes of data.

Moreover, with the emergence of Web 2.0, a new vision of the Web put the user at the center of information considering it as a potential producer of web content and not just a consumer [8]. This radical change has significantly increased the amount of information [10]. Consequently, the proliferation of types of information from multiple sources such as social networking, services, blogs, information aggregation websites, videos, images, text, creates a wide variety of data types beyond traditional relational data. These data do not exist in a perfectly ordered form and are not amenable to analytical operations. They no longer fall within the net structures, easy to consume, rather they are semi-structured or unstructured. This is within the aspect range of data which represent the second V in the design of the Big Data phenomenon. And consequently, this leads to ask: how it is possible to deal with and process these unstructured data to make them consumable by human users and / or applications?

Indeed, it is possible to find this problematic of data variety on educational purposes, and especially during a online search or an e-learning process. The learner that uses the e-learning platform is left with a mix of data that do not meet necessarily their expectations and they sometimes even prove useless against expected results. It is in this context that it is proposed to develop a methodology based on a tool called "Big-Learn" to integrate the mix of structured and unstructured data in one data layer to facilitate access and more optimal relevance search with adequate and consistent results according to the expectations of the learner. So, is there a fictitious model to represent and process this type of data that are not necessarily text? Also, is the semi-structured databases NoSQL provide enough structure to organize the unstructured data? knowing they do not require an exact schema data before storing.

The following paragraph present the state of the art concerning Big Data in the e-learning environment and the describtion of the problematic of data variety and their impact on the educational purposes ; the paragraph 3 expose the concepts and approach of the Big-Learn tool using the online search as case study. The last paragraph present a general conclusion putting forward a series of perspectives.

II. Big Data in E-Learning Environment

A. Definition and state of the art

The society has experienced in recent years the arrival of Big Data phenomenon, So much data is available and requires powerful computers and algorithms process them. Attal Butte estimates the volume of data produced each year to four-zetta bytes (1 byte = zetta-10 to the power 21 bytes) [1]. Currently, 2.5 terabytes of data are produced every day in the world. By the year 2020 it is estimated that the data size will be multiplied by 50. Google receives 40,000 requests for information every second, 72 movies are set to YouTube every minute and 217 new Smartphone users are counted every minute. [7] Today the information is coming from all sides: geolocation sensors, data from smartphones (connection logs, appeals, etc.), data posted on social networks, video and satellite images, Transactions customers, sensor forms of movement or connected objects, etc. Concretely, this is the real-time development of a large mass of data that goes far beyond the capacity of conventional processing and analysis

tools (relational databases, SQL, etc.). These mass data needs to be analyzed and processed for their use and consumption on the part of users and applications.

However, a number of authors [2] [6] postulate some provocations on the Big Data phenomenon that require critical thinking. The increasing use of big data questions some assumptions about traditional knowledge in the context of a claim of Big Data to "objectivity and accuracy [who] are deceiving." It is also necessary to realize that "all the data is not equivalent" [2]. To better explain this, the Table 1 summarize some advantages and limits of Big Data phenomenon.

TABLE I. SOME ADVANTAGES AND LIMITS OF BIG DATA

Advantages	Limits
• The ability to search and cross massive data sets; • Completeness of the perimeter and the capture of entire populations of the systems; • Targeting maximum detail, aiming the common fields for the combination of data sets; • The flexibility and extension: Easily add new fields; • Scalability: the potential to grow rapidly; • The prediction of future performance and identify potential problems; • The interpretation of the actual operational data; • The evaluation of the performance of the organization or institution; • Decision making thanks to the researched information at the right time.	• The creation of new fractures: increasing inequalities and injustices that exist. • The complexity of managing and analyzing large amount of heterogeneous data. • Uncertainty of informations coming from different sources; • The loss of value and / or reliability of data coming from different systems. • Problems related to the ethics and confidentiality of information; • Lack of regulations against abuse and bad data uses. • Misinterpretation of information; • Collecting incorrect data that can lead to erroneous results.

As it is shown, the Big Data phenomenon offers several benefits for the completeness and flexibility of data use that themselves can cause bad consequences on the objectivity and accuracy of the interpreted information. The analysis and exploration of big data exceeds human capacity, which requires the use of computer systems powerful and able to explore them. But these data does not occur in an ordered form and are not ready for analytical operations, they rather come in a semi-structured or unstructured form. It is this dimension that concern the subject of this article and especially the problematic of the non structured types of data produced by Big Data.

The Big Data phenomenon has obviously impacted the learning environment and the distance training. It has facilitated the creation of a mixture full of learning opportunities and allows the learners to improve their training practices and experiment with open educational resources, especially the massive and open online courses (MOOC) and the distance learning via e-learning platforms. With the emerging technologies in the Web, access to information has become easier with the ability to work and learn effectively, regardless of educational structures that have been the norm for centuries [3] [4]. This put in place new structures and new working environments, enabling independent learning, but that does not mean that everyone is able to do it effectively [5]. Two major factors are the basis for the study of learning in a massive and open online environment: learner's autonomy and the quality of the massive submitted information. It is this last factor that concern the subject of this article.

B. Problematic of variety and non-structuring massive data and their impact on the educational purposes

Big Data knows several challenges and opportunities as well as involvement of several technologies because of the flood of data produced each year by users and companies. The information of the Big Data comes from many sources of data. The Table 2 shows some examples of these sources like the Web, the Internet, communication objects, genomic sciences, astronomy, commerce and public data.

TABLE II. SOME DATA SOURCES OF BIG DATA

Data source	Examples
Web	Access logs, social networks, e-commerce, documents, photos, videos, etc. (example: Google treated 24 petabytes of data per day with MapReduce in 2009).
Internet and communication objects	RFID, sensor networks, telephony call logs.
Genomic science, astronomy, subatomic physics, climatology, etc.	CERN announces generate 15 petabytes of data per year with the LHC. The German research center on climate manages 60 petabytes of database.
Marketing	The transaction history in a hypermarket chain.
Public	Open Data

As presented in the table, the integration of data in the Big Data processes covered several unstructured data (sensor data, web logs, RFID, social networks, documents).

It is possible to find the problem of the data variety for educational purposes especially in online search or distance training. The learner that uses the e-learning platform is left with a mixture of data such as files of all kinds, curves, course materials, quizzes, etc. Also, in the educational component, the social tools of web 2.0 allows to create and publish any type of educational content such as lectures, exercises, assignments or bibliographies and digital resources enable an informal collaborative learning known as "Learning 2.0"[9]. However, these data do not always meet the needs of the learner and they sometimes even prove useless against expected results. A new challenge is then born in data analysis, it is to make significant progress in the process of this type of unstructured data which the amount is continuously growing. To do that, the use of specific IT tools for the processing of unstructured data has become essential. Knowing that relational databases are not always the best solution because of their static patterns.

III. TOWARDS A BIG DATA SYSTEM FOR PROCESSING MASSIVE DATA IN THE E-LEARNING ENVIRONMENT

A. Case study: the use of the online search

1) Context

Since becoming aware of innovations in learning, technologists have begun to design and develop tools to help

learners to better understand this new way of teaching connected to data in constant evolution. To success the purpose of the e-learning, it is crucial to create a trusted environment where learners feel comfortable. A place that can aggregate content and imagine it as a community where dialogue flows and interactions and content can be simple to use. This will enable learners to develop clear ideas and evolve in their learning in depth.

In this context, it is proposed to create an educational platform that would support learners in their environment. Research involving the design and development of this platform is working in many directions, but here the object is to report some progress in education, advances on issues concerning self-learning and online search in a massive data environment. And to better understand the study, the choosen case is the use of the online search by learners who seek to acquire information about a specific course or theme given for the purpose of learning and documentation, and, to see if there are other extra dimensions that could be added following the study on learning in a massive data environment.

The results of the data analysis for the online search scenario will allow to delineate the context of the future system and to better understand the design of a methodology based on the tool that will integrate the mixing of structured and unstructured data in one layer to facilitate access in addition to an optimal search relevance with adequate and consistent results according to the needs of the learner. The Fig. 1 shows the Big-Learn system usage scenario illustrated by a sequence diagram.

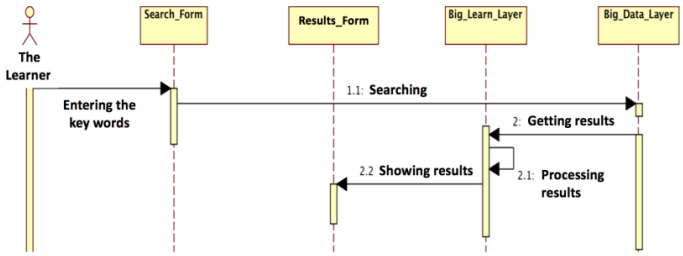

Fig. 1. The sequence diagram of the general use for the Big-Learn system

The user of the system (the learner) accesses the search interface to enter the keywords of his information request. These keywords are sent to the system for retrieving information corresponding to their semantics from the Big Data layer. Then, the results data is processed at the Big-learn layer to structure, classify and send the processed data to the results page of the system, which will be displayed to the learner, the user of the Big-Learn system.

2) Methodology

To a better design and development of the futur system, it is necessary to study the circumstances of the learning that takes place on online networks and distance training. It is important to know the relevance of the learning experience of people in online networks in which they find the information they are likely to consume. As part of this study, informational or learning data are defined as data collected from online open spaces where people access remotely, while communicating with others via blogs, audiovisual, wikis and other sources of information and other remote communication resources. Constraints and challenges emanating from such an environment show themselves well in problems related to the study of human behavior, as well as other constraints involved, including the variability of the network and data, power relations on the network size and the generated content. Relevant analysis requires an approach by mixed methods and results in a new ethics and questions about the confidentiality of information or data.

For this study, it is proposed to conduct a sampling survey concerning the criteria of the future system and indicators of performance and learning quality in the use of online search, through the design and the submission of survey to a set of students who represent the future users of the online search and distance training platform. The content of the survey describe the different factors suceptibles to impact and affect the learner when using the online search, and the criteria of simplicity and ergonomics of the use of the future solution.

The questions of the survey were raised in relation to the aspect of the presentation and quality of information in the online search results in a documentation or on a distance training course or theme. The classification of information also plays a dominant role in the organization and fair presentation of the results of the search. The relevancy of the result and the content is crucial to the learner using online search. Thus, several elements must be examined and redefined to design the right solution.

The Google Forms tool will be used to design and submit the survey. It is a tool to plan events, to conduct a survey or poll to subject students to a questionnaire or to easily collect information online. Forms can be created from Google Drive or an existing spreadsheet can collect answers to questions on the form. The target audience consists of two samples, the first few groups of students from the ESI (Science School of Information) of Rabat, and the second is a group of students from the FS (Faculty of Science) of Tetouan.

The goal of this investigation is to identify the key elements and factors that may impact and influence the environment when the learner uses the online search for learning purposes and documentation on a given theme or topic, and to enable to deduct thereafter performance and functioning indicators related to learning in the Big Data environment to better design and implement an adequate system and meeting the expectations of most learners. Table 3 highlights some examples of factors to consider elements during this investigation.

TABLE III. THE FACTORS ELEMENTS OF THE ONLINE SEARCH SURVEY

Data source	Examples
The type of search result that the learner prioritizes at its online search	video, document, image, article, ...
The criteria of a relevant and fractueuse search	Number of views, date of publication, source of the element, ...
The number of results that the learner prefer by page	5, 10, 50, ...
The response time in seconds that the search query take to be sent	1s, 5s, 10s, ...

As described in this table, several factors can influence the process of using the online search by the learner. Thus, the results of its application will depend on the relevance and proper configuration of these criteria.

B. The Big-Learn System

It is important to identify a sophisticated strategy to combine different types of data in a way that they provide the best result to the learner, the user of the e-learning platform. In this context, it is proposed to develop a technique based on a tool called "Big-Learn" that integrates the mix of structured and unstructured data in one data layer to facilitate access in addition to an optimal relevance of search with adequate and consistent results according to the expectations of the learner. The adopted method will consist initially in a quantitative and a qualitative study of the variety of data and their typology, followed by a detailed analysis of the structure and harmonization of the data to finally find a fictional model for treatment of such data. This conceptual work will be crowned with a working prototype as a tool achieved with UML and Java technology. Fig. 2 shows the functional architecture of the Big-Learn system.

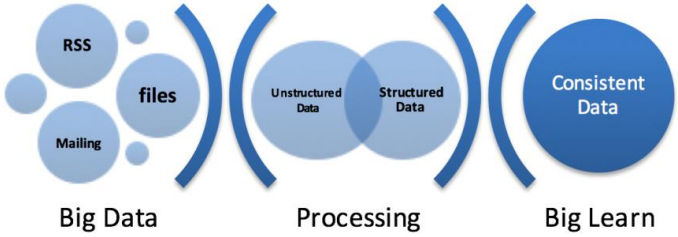

Fig. 2. The functional architecture of the Big-Learn system

As is described, it is proposed to develop a top layer to the raw and varied data coming from "Big Data" providing thereafter the storage, retrieval and dissemination of consistent information to research information requested by the user of the Big-Learn platform.

The scenario of the use of the online search that has been already mentioned above, can be achieved through the application components of the Big-Learn system as described in Fig. 3.

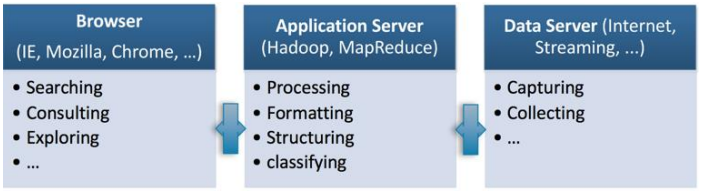

Fig. 3. The application architecture of the Big-Learn system

Thus, the data server will handle the capture and collection of massive data (Big Data) and then, the application server can carry out the treatment, structuring, formatting and classification required for these raw data and thus make them consumable by the presentation layer at the end and that will be accessible via the web browser by the user of the Big-Learn system.

The interface of the tool will be an easy space to use for the learner to make online search for learning or documentation on a specific subject or theme, the example shown in Fig. 4 shows the use of the system via its interface to search information about the subject of the use of the system "Viber": The application for free calls and messages.

Fig. 4. The user interface of the Big-Learn system

After entering the keywords to search the specific thematic as shown in the previous figure, the Big-Learn system performs the capture of all types of data (text, image, video, audio, etc.) related to the subject of the theme and group them in its raw data layer as shown in Fig. 5. It then includes data of any type, such as posts, pictures, videos, audio tracks, etc.

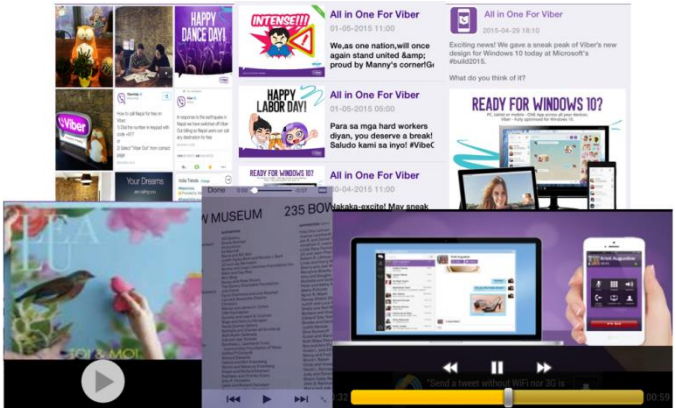

Fig. 5. The raw data collected related to the requested theme

The big-Learn system thereafter proceed with the treatment of the raw data to make it consumable by the user, via classification, structuring and formatting of these data at the presentation layer, thus allowing an organized display and ergonomics at the user interface of the system, as illustrated at Fig. 6.

Fig. 6. The organization of results data according to their typology

As it shown, the data related to the video type classified in the same category even though they come from different

sources (youtube, vimeo, dailymotion, ..), also, for the type of post data (facebook, twitter, google +, ...), audio data type (soundcloud, deezer, ...) and finally the type of the image data (instagram, picassa ...).

IV. Conclusion and Future Work

The available tools actually on the market make it possible either to analyze structured or unstructured data, but not both at the same time. Consequently, little Big Data technologies provide integration of various types of data and align with the structured data. It is in this context that it is proposed to develop a methodology based on a tool to integrate the mix of structured and unstructured data in one data layer to facilitate access and more optimal relevant search with adequate and consistent results that meets the expectations of the learner. The solution will also enable the detection of language elements, turn them into a data type that can be manipulated and be the object of the processing of the consumption of the information. The adopted method will consist initially in the study of the criteria and factors impacting the environment of the learner towards massive data offered by the Big Data via the case study of using the online search for learning purposes or documentation on a given theme. This, through the creation and submission of the survey corresponding to a sample of learners using online search for their learning. This will be followed by a detailed analysis of the results collected from these survey and that will frame the functional and technical requirements of the future solution to finally design a hypothetical model for the treatment of these heterogeneous mass of data.

References

[1] Attal Butte. Big Data, Big machines, Big Science : vers une société sans sujet et sans causalité ? (2014). Adult Education, 2014, vol. 26, no 1, p. 35-55.

[2] Boyd, D., & Crawford, K. (2013). Six provocations for Big Data.

[3] Downes, S. (2010, May 12). The role of the educator.

[4] Fournier, H., & Kop, R. (2011) Factors affecting the design and development of a Personal Learning Environment: Research on super-users, in the International Journal of Virtual and Personal Learning Environments, 2 (4), 12-22.

[5] Kop et Bouchard, 2011 Kop, R., & Bouchard, P. (2011). The role of adult educators in the age of social media. In M. Thomas (Ed.), Digital education:Opportunities for social collaboration (pp. 61-80). New York, NY:Palgrave Macmillan.

[6] Lyon, D. (2014). Surveillance, Snowden, and Big Data: Capacities, consequences, critique. *Big Data & Society*, July-September pp. 1—13. DOI: 10.1177/20253951714541861.

[7] Miranda, S. (2013). « De Big Brother au Big Data », Conférence de Big Data, Université Sophia Antipolis.

[8] O'Reilly, T. 2005. "What Is Web 2.0, Design Patterns and Business Models for the Next Generation of Software",

[9] SBIHI, Boubker, EL KADIRI, Kamal, et AKNIN, Noura. Towards an Implementation of the Concepts of E-Learning 2.5 through one Group of ten Master's Learners: Case of the UML Course. *International Journal of Emerging Technologies in Learning (iJET)*, 2013, vol. 8, no 4, p. 68-73.

[10] SBIHI, Boubker et KADIRI, Kamal Eddine El. Towards a participatory E-learning 2.0 A new E-learning focused on learners and validation of the content. *arXiv preprint arXiv:1001.4738*, 2010.