Processing big data with natural semantics and natural language understanding using brain-like approach

Emdad Khan

Abstract – We propose semantics and associated Natural Language Understanding (NLU) based approach to address the key problems of big data. Our approach use human Brain-Like and Brain-Inspired algorithms as humans can very effectively retrieve knowledge from data as well as can significantly compress the data by using the semantics of the information. This is true for both unstructured and structured data which are growing very fast - already exceeding the exabyte range. Unstructured data, however, dominates structured data with a wide margin.

There are multiple problems with big data including storage, search, transfer, sharing, analysis, processing, viewing, and deriving meaning / semantics. Such problems are mainly due to the 4 Vs i.e. Volume, Velocity, Variety and Variability. All these problems can be addressed well when the data can be highly compressed and meaning of the data can be converted to knowledge. We propose to use Semantic Engine using Brain-Like Approach (SEBLA) to convert data to knowledge and also to compress it; thus addresses the Big Data problems in an effective way. SEBLA provides "Natural Semantics" i.e. semantics similar to what humans use.

The main theme in SEBLA is to use each word as object with all important features, most importantly the semantics. In our human natural language based communication, we understand the meaning of every word even when it is standalone i.e. without any context. Sometimes a word may have multiple meanings which get resolved with the context in a sentence. The next main theme is to use the semantics of each word to develop the meaning of a sentence as we do in our natural language understanding as human. Similarly, the semantics of sentences are used to derive the semantics or meaning of a paragraph. The 3rd main theme is to use natural semantics as opposed to existing "mechanical semantics" of Predicate logic or Ontology or the like. All these together can achieve summarization and draw inference i.e. finding useful information or converting data to knowledge, at least to first degree. These have numerous applications (e.g. Business Intelligence, Analytics, Document Summarization, and Document Analysis). There are other types of key applications using the semantic capability of SEBLA including Intelligent Information Retrieval, Intelligent Search, Question and Answer System, and Intelligent Language Translation. We have presented a few such applications.

Emdad Khan is with the College of Computer & Information Science, Imam University, P.O. BOX 5702, Riyadh , Saudi Arabia. He is also with InternetSpeech, Inc, San Jose, CA, USA (Phone: 408-532-9630, fax: 408-274-8151, email: emdad@internetspeech.com)

***Keywords* ----** Big Data, Unstructured Data; Natural Language Understanding (NLU); Semantics; Artificial Intelligence; Data Mining; Knowledge Discovery; Data Science; Interpreting Data; Internet; Intelligent Internet; Question & Answer System; Intelligent Agent; Machine Learning; Predictive Analysis; Business Intelligence; High Value Business Problems; Information Technology.

I. INTRODUCTION

The amount of data in our world has been exploding, and using (analyzing, processing, searching, storing and understanding) large data sets - so called big data - has become a critical issue providing both challenges and opportunities. The increasing volume and details of information captured by enterprises, the rise of multimedia, social media, and the Internet of Things will fuel exponential growth in data for the foreseeable future [1].

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data [4]. Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics, proteomics, metabolomics, economics, complex physics simulations and biological & environmental research

Big data usually include data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target - as of 2012 ranging from a few dozen terabytes to many petabytes of data in a single data set. The target moves due to constant improvement in traditional DBMS technology as well as new databases like NoSQL and their ability to handle larger amounts of data. With this difficulty, new platforms of "big data" tools are being developed to handle various aspects of large quantities of data.

One key area is handling of unstructured data. This is usually a high value business problem as it can save (or generate) significant amount of money by doing good Predictive Analysis, and hence providing a good Business Intelligence (BI).

Handling of large unstructured data is also very important from several other aspects including Intelligent Information

Retrieval, Intelligent Search (getting more relevant information), more accurate Language Translation and Question & Answer (Q & A) System.

The Big Data industry is growing fast. In 2010, this industry on its own was worth more than $100 billion and was growing at almost 10 percent a year: about twice as fast as the software business as a whole ([1], [4]).

Many scientists, engineers, researchers and others have been working on Big Data. They have been using various algorithms, methods and technologies including A/B testing, association rule learning, classification, cluster analysis, crowd sourcing, data fusion and integration, ensemble learning, genetic algorithms, machine learning, natural language processing, neural networks, pattern recognition, anomaly detection, predictive modeling, regression, sentiment analysis, signal processing, supervised and unsupervised learning, simulation, time series analysis and visualization [4].

While existing solutions are very valuable, many Big Data problems are not solved / not solved efficiently yet. E.g. summarization of large documents, interpreting data, drawing inference or knowledge discovery, and prediction (especially for structured Big Data). There are plenty of skeptics when it comes to computers and algorithms abilities to predict the future. People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately. People's environments change even more quickly than they themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the exact same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions [14]. This is probably an exaggeration as we know statistical techniques have been successful for prediction problems but with limitations, e.g. Bonferroni Correction [18]. Other existing techniques also have limitations, although, in general, produce better prediction. In fact, "making good prediction" is still an open problem.

Solving Big Data problems will become more and more important and challenging as data size grow, a trend that will continue for a long time. At the same time, this also brings more opportunities in various ways as reflected by the growth of Big Data industry as already mentioned.

In this paper we have addressed the Big Data problems for both unstructured and structured (with more emphasis on unstructured) data using **Semantics and Natural Language Understanding (NLU)**. As we know, humans do a very good job in processing unstructured data, especially using semantics to significantly compress the data by representing with a few words or sentences using the semantics of the information while keeping the core meaning. Thus, semantics and NLU help human for knowledge discovery / drawing inference, deriving intelligence, doing summarization as well as significantly compressing the data. Semantics also plays a very important role in processing and deriving knowledge & intelligence from structured data.

Since our approach is based on the way we believe our brain works (as also evidenced by many researchers), we call it **Semantic Engine using Brain-Like approach (SEBLA)** even though we do not know exactly how our brain does it.

Section II describes the nature of data growth and "Data Equity". Section III provides key problems of Big Data. Section IV describes key issues in dealing with unstructured and structured Big Data. Section V describes our approach using SEBLA to handle big data. Section VI describes how our approach can be used on structured data. Section VII describes how our solution can be used in various applications, Section VIIII describes Future works and Section IX provides conclusions.

II. NATURE OF DATA GROWTH AND DATA EQUITY

In this Information Age, information is growing very fast. Internet is a classic example. The data growth on the Internet during last 15 years is phenomenal. Internet has changed the world as we all know. There are various other key sources for data growth – e.g. scanners, sensors, mobile phones, smart meters, social media platforms, credit cards, digital medical records, satellite imagery and the like. Such data sources generate both unstructured and unstructured data. Such data are also getting integrated on the Internet and Intranet. With the growth of data, the nature of its usage is changing fast. E.g. the field of astronomy is changing from where taking pictures of the sky was a large part of an astronomer's job to one where the pictures are all in a database already and the astronomer's task is to find interesting objects and phenomena in the database. In the biological sciences, there is now well established tradition of depositing scientific data into a public repository, and also creating public database to be used by other scientists [19]. In the business world, Business Intelligence (BI) has already become an important field to extract key business data from large volume of data.

Like "time is money", "information is money" in this information age. Accordingly, there is high value in Big Data as it has lots of information. Question is how to find valuable data from sea of data and monetize that in an efficient way. The advent of Big Data Analytics in recent years has made it easier to capitalize on the wealth of historic and real-time data generated through supply chains, production process and customer's behaviors. The value of Big Data that can be unlocked from analytics is also known as **"Data Equity"** and it is increasing rapidly [20].

Many people have started putting great efforts to bank on the Data Equity. Such efforts have resulted various Big Data Analytics tools - for example, predictive analytics, data mining on Big Data, and advanced data visualization. While these are good, much more works are needed to maximize the Data Equity. We would need to address key Big Data problems for both unstructured and structured data, especially, in interpreting, summarizing and drawing inference / knowledge discovery. Apart from analytics, there is Data Equity by solving all other Big Data and associated problems. For examples, just doing a “semantic search” or “semantic data compression” (see Section 4) can provide high Data equity.

III. KEY PROBLEMS OF BIG DATA

There are multiple problems with big data including storage, search, transfer, sharing, analysis, processing, communicating, viewing, extracting value, deriving meaning / semantics, knowledge discovery and decision making. Such problems are mainly due to the 4 Vs i.e. Volume, Velocity, Variety and Variability. A few other issues are security, privacy and usability. Velocity means the rate data arrives and the time it needs to be acted upon. Variety means the heterogeneity of data types, representation and semantic interpretation. Variability means the variation in the data.

The analysis (modeling) of Big Data involves multiple distinct phases as shown in Fig. 1 (courtesy [19]) each of which introduce

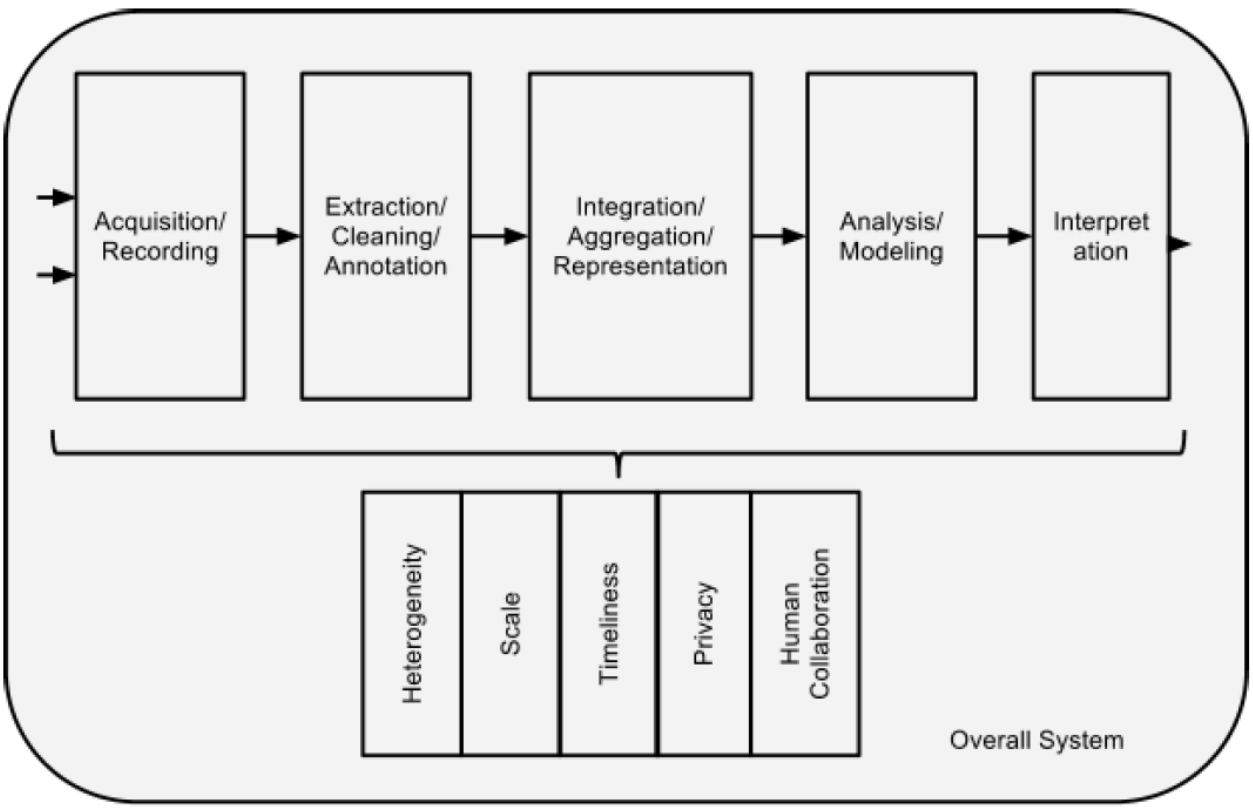

Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

challenges. While analysis of Big Data is very important, other phases of the data analysis pipeline are also very important. These other phases, especially, interpreting data and discovering knowledge from it are also very complex.

Data visualization and dissemination (not included in Fig. 1) are also important and complex steps. As we know picture is worth thousand words – especially when we are trying to understand and gain insights from data. It is particularly important when we are trying to find relationships among thousands or millions of variables and determine their relative importance. Such insights are critical for several things including development of models or algorithms.

IV. ISSUES DEALING WITH UNSTRUCTURED BIG DATA

Unstructured data dominates the data world. It is estimated that over 80% data in computers and Internet are unstructured [5]. Unstructured data can be broadly classified into two groups:

(a) Text data and
(b) Non-text data (including sound, image video).

Computers are very good in processing structured data (e.g. data in a database). This is mainly because computers are still mathematical devices, especially, fast number crunchers. When it comes to unstructured data, we are dealing with the meaning or semantics and associated context; and humans are very good at that.

In the textual case there is a key problem of context. The classic example often given is the difference between the statements that "John rides in a mustang" and "John rides on a mustang" [5].

A human analyst will see a great deal of difference between these two sentences. Our experience adds enormously to our understanding of both. We know, for a start, that the first statement refers to a car, the second to a horse. But we will also understand that in the first statement John is a man, and he is probably in the United States, because Ford Mustangs are not sold in large numbers outside the US.

In the second statement, we may consider that the event might have occurred in the US as the descriptive term is generally associated with that country, and a long time ago, as there are not many wild horses left in the US. It might even occur to us reading one of the two sentences that, because the O and I keys are beside one another on a standard keyboard, there could have been a typographical error and the other sentence may be the correct one.

The human brain picks up all of this data almost instantaneously – our understanding is implicit. Computers cannot deal with implicit information and have to be told how to understand it. Consequently they deal with this 'tacit' information very badly, if at all.

This gets further complicated as the writing style, sentence structure and vocabulary used in formal documents are very different to those used in e-mails, which are in turn different to those used in text messaging. Humans can handle all these very well.

One classical approach used in a computer to handle text data is "keyword" or key phrase search. Although useful, this method is far from perfect. If the set of search terms is too narrow it can miss vital information, if it is too broad the resulting set of 'hits' can contain large numbers of totally irrelevant 'false positives'.

Modern search tools have improved things somewhat. Computerized thesauruses allow us to search for synonyms and homonyms without having to explicitly set out every possible variation. Other tools allow for 'stemming' - for example, in Lexis Nexis putting in the term 'run+' will cause the engine to search for 'run', 'runs', 'running', 'runner', and so on.

One key modern method is the use of some semantics using Predicate logic, Ontology and the like. However, one would need to define clearly all such semantics. Any small variation in the words or structure can cause the semantics to be different yielding wrong results or no results. Such approaches basically provide some "mechanical" semantics; thus limiting them to applications with small domain.

The problem becomes even more critical when we try to use non-text data – like audio, image, video. Here also, human brain handles such data very efficiently.

Thus, existing approaches have simplified the process somewhat, but they still have not solved the problem of computers' inability to deal with tacit and context-based information. At present, we can conclude that text analysis technology may be better at data reduction than actual data analysis. As already explained, human brain is very good in addressing these problems. The key point is that we would need to use the semantics and NLU capabilities in dealing with unstructured data (see Section V).

It is important to note that although humans can do text processing very well, they can do it only for relatively smaller size data. Human brain cannot handle very large data like big data. However, using human brain's intelligent approach with the fast number crunching computers, we believe, we can effectively solve the big data problems - the theme of our approach.

V. SEMANTICS AND NLU TO HANDLE UNSTRUCTURED BIG DATA PROBLEMS

The key problems associated with unstructured data (as described in the previous Section) are related to the semantics of words, sentences and paragraphs. As mentioned, human brain uses semantics and natural language understanding (NLU) to very efficiently use unstructured data. Below, first we briefly describe a Semantic Engine ([6], [7], [8]) using Brain-Like algorithms (SEBLA). Then we show how SEBLA can handle Big Data.

5.1 **Semantic Engine - SEBLA**

While traditional approaches to NLU have been applied over the past 50 years and had some good successes mainly in a small domain, results show insignificant advancement, in general, and NLU remains a complex open problem. NLU complexity is mainly related to **semantics**: abstraction, representation, real meaning, and computational complexity. We argue that while existing approaches are great in solving some specific problems, they do not seem to address key Natural Language problems in a practical and natural way. In [8], we proposed a Semantic Engine using **Brain-Like approach (SEBLA)** that uses Brain-Like algorithms to solve the key NLU problem (i.e. the semantic problem) as well as its sub-problems.

The main theme of our approach in SEBLA is to use each word as object with all important features, most importantly the semantics. In our human natural language based communication, we understand the meaning of every word even when it is standalone i.e. without any context. Sometimes a word may have multiple meanings which get resolved with the context in a sentence. The next main theme is to use the semantics of each word to develop the meaning of a sentence as we do in our natural language understanding as human. Similarly, the semantics of sentences are used to derive the semantics or meaning of a paragraph. The 3rd main theme is to use natural semantics as opposed to existing "mechanical semantics" of Predicate logic or Ontology or the like.

A SEBLA based NLU system is able to:

1. Paraphrase an input text.
2. Translate the text into another language.
3. Answer questions about the content of the text.
4. Draw inferences from the text.

As an example, consider the following sentence:

"Maharani serves vegetarian food."

Semantics represented by existing methods, e.g. Predicate Logic, is

Serves(Maharani, Vegetarian Food) and
Restaurant(Maharani)

Now, if we ask
"is vegetarian dishes served at Maharani?"

the system will not be able to answer correctly unless we also define a semantics for "Vegetarian Dish" or define that "food" is same as "dish" etc. This means, almost everything would need to be clearly defined (which is what is best described by "mechanical semantics"). But with SEBLA based NLU, the answer for the above question will be "Yes" without adding any special semantics for "Vegetarian Dish".

The "mechanical semantics" nature becomes more prominent when we use more complex predicates e.g. when we use universal and existential quantifies, and/or add constructs to represent time.

It is important to note that ML (Maximum Likelihood) based performance commonly used in prediction (e.g. when one types words in a search field on a search engine it shows the next word(s) automatically) will be improved with natural semantics. Currently, mainly ML (and sometimes other techniques including existing semantics methods) is used for prediction. By using proposed more natural semantics (e.g. using SEBLA), the meaning of the typed words will be more clear; thus helping better prediction of the next word(s). It will also help using natural sentences in the search field than special word combinations, e.g. when using advanced search.

5.2 Using SEBLA to Handle Unstructured Big Data

To handle unstructured Big Data, an Intelligent Agent (IA) is used that utilizes semantics of SEBLA and NLU in various ways depending on the task. The Big Data tasks can be broadly classified as:

a. Information Retrieval (IR) / Search

b. Question & Answer System

c. Summarization

d. Converting data to information to knowledge to intelligence

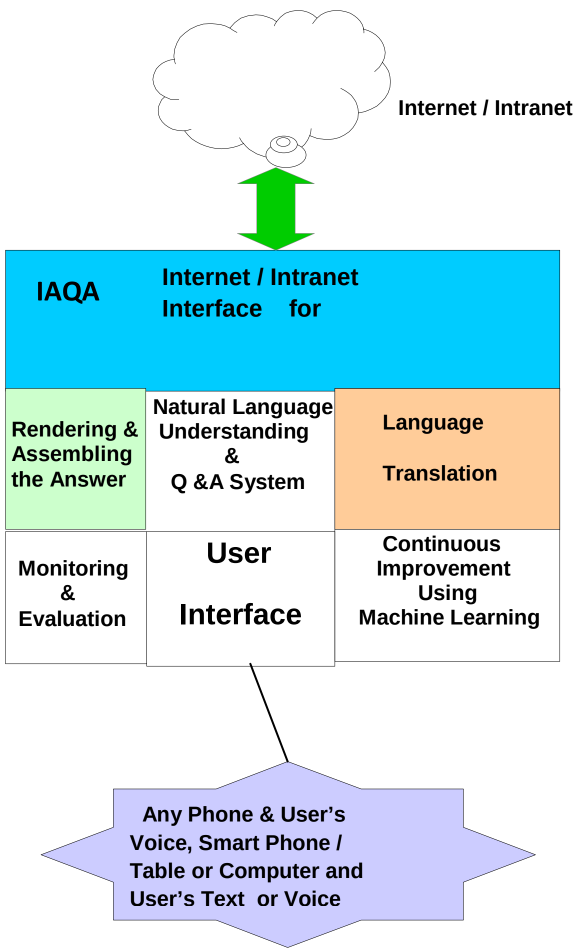

Fig. 2 IAQA: Intelligent Agent for a Question & Answer (Q & A) System.

Note that all these do significant data compression that helps other key features of Big Data including storage, processing, and visualizing. E.g. in IR, instead of retrieving all information using string search, SEBLA will reject all information that is not related semantically i.e. it will retrieve information that are related semantically. Similarly, for data mining, only semantically meaningful data will be found which can be much smaller than the results using existing mining methods.

For the key tasks of IA, let's consider the case of a Q & A System. The key tasks for this case are:

1. Understand user's request and break it into key component parts.
2. Act on all the component parts, find requested answers by accessing appropriate sources (including database tables).
3. Assemble a concise answer, and then present it in a nice way.

The IA itself also uses SEBLA's natural semantic engine to make correct decisions by avoiding "mechanical semantics", as commonly used in existing systems. Such an IA for Q & A system (IAQA) is shown in Fig. 2. The term "**rendering**" ([9], [10], [11]) needs some explanation. As we know, the Internet was designed with visual access in a relatively large display screen (like a 8.5 inch x 11 inch page) in mind. Thus, all the content are laid out on any website and webpage in a manner that attract our eyes in a large screen. Retrieving the desired content (which is much smaller in size than the total content on a webpage or website) from a typical webpage / website and displaying that (or playing in audio) into a much smaller screen (like in a cell phone or PDA) is a very challenging task. This process of retrieving and converting most desired content from a large source of content into a much smaller but desired content is called **"rendering".** Clearly, rendering is mainly related to Internet Browsing on a small device. A Q & A system uses rendering to get an initial answer and then further refines it with semantics. Rendering includes form rendering, retrieving appropriate data when a form is submitted, and retrieving multi-media data. A Q & A system also uses rendering to get appropriate data from various websites, via web services and other query methods.

VI. SEMNATICS AND NLU TO HANDLE LARGE STRUCTURED DATA

Structured data are much smaller in size compared to unstructured data and computers can handle structured data well. Thus, it may appear that the need to more efficiently address structured data is not that critical. While this perception is partially true, the need to more efficiently address structured data is also very important. The key reasons are:

a. structured data are also growing very fast and conventional algorithms are not enough in many cases.
b. many Big Data applications, although are dominated by unstructured data still needs structured data (e.g. analysis report in a BI)
c. meanings of structured data (e.g. relationships between many variables) are critical to process them effectively and efficiently.
d. all key steps in the analysis pipeline (Fig. 1 including visualization) are very important to understand and interpret structured data.

Thus, most of the issues related to unstructured data are also equally applicable for structured data. Accordingly, semantics and NLU can be efficiently applied for structured data.

Let's take an example of relationships between various data fields in various tables in a database. Today's approach using database programming (e.g. using a set of SQL queries and some associated conclusions) becomes difficult when relationship size and data size grow. Besides, such relationships are defined "mechanically" sometimes using "mechanical semantics" as explained for unstructured data.

In contrast, let's consider that data table headings have natural words or sentences. Using the semantics of such words or sentences, it would be much easier to express such relationships. Moreover, semantics will enable to define many complex relationships that cannot be defined currently. Via appropriate data-mining & other techniques and the use of semantics, a significant data compression will also be possible.

Another key contribution is database User Interface via NLU. A NLU based interface can take natural sentences and automatically convert them into equivalent database queries and then assemble the retrieved information in a simple usable form. This will make the implementation of various Big Data applications (including a Q & A) simpler and more useful. Today, using a relational (or other) database is not easy and intuitive for many people, especially those who are

a. New to interface with a database for general use (e.g. a manager or a business intelligence expert). Such people use some programmers to develop some simple interface to a database with limited commands.

b. New to software programming and new to learn how to program with database.

Database programming is still very important but semantics and NLU based interface will make the process more efficient and useful, especially, in addressing the problems related to structured Big Data.

VII. SAMPLE APPLICATIONS

Our SEBLA and NLU based approach can be used in various applications including Intelligent Information Retrieval, Intelligent Search, Q & A System, Summarization, and Business Intelligence. Below we have provided a brief description of two example applications:

(1) An Intelligent Information Retrieval (IIR) system.

(2) An Intelligent Summarization system.

7.1 An Intelligent Information Retrieval (IIR) system

The information retrieval through existing IR and search engines are mainly based on string search. Thus, the search process needs to deal with many data to find matches. And all matched data are extracted even though many data are not relevant and desired. Accordingly, such engines produce many (often thousands of) results, and human knowledge and intelligence are needed to retrieve the desired information from such search results. This requirement usually limits the usage of search engines to experienced and educated users. There are FOUR key issues with the current approaches:

a. Search process needs to deal with very large data.
b. String search results contain many undesired and unrelated results.
c. String search results may not contain the desired results and user may need to do multiple searches by various search word combinations.
d. String search results may NOT contain the desired information even after trying major key word combinations as a user may skip key words of similar meaning.

The semantic capability of SEBLA addresses these issues in TWO broad ways:

A. Retrieve expanded and more related information and then get most desired information by filtering.

B. Retrieve far less but more related and appropriate information and then get more refined desired information.

Approach in A (an IA similar to Fig. 2 but instead of Q & A, it does retrieve key information) is useful when string search data is not too large and conventional search engines can be used. The key steps using approach A are:

1. In the query sentence / string to understand the meaning of each word and sentence.

2. Generate all related sets of query strings using semantic meaning of each word and sentence (thus generating lot more appropriate search results that are related to the input words and sentences).

3. Extract the most appropriate and related results from the extended search results. This is achieved by employing the semantics and rendering (Fig.3 – corresponding to the "rendering & assembling the answer" block in Fig 2).

Many words have multiple synonyms. By understanding the semantics of each word, a complete (or nearly complete) set of synonyms will be generated. Without semantic meaning, only limited predefined synonyms can be used as done for some words in existing search engines. This is also true for sentences. By understanding each search sentence, corresponding equivalent search sentences and corresponding words will be generated. The sentence level semantics will be used to refine the word list to help reduce search results when submitted to search engines.

If a user only presents search key words and no sentences, then, using NLU, a set of most relevant search words will be generated. This is be done by creating different word combinations (including all synonyms), deriving the semantic meanings, and then appropriate filtering to derive the most appropriate set of search word combinations.

If a user presents sentences, then similar sentences using synonym words will be generated by keeping the context same. Then corresponding set of key words will be generated. This is important as existing search engines mainly work on string search and does not depend on the meaning of the sentences or words. However, they do strongly consider word combinations.

Approach in B is useful when string search data are very large and conventional search engines can take too long. This approach is more appropriate for Big Data. The key steps using this approach are:

1. In the query sentence / string to understand the meaning of each word and sentence.

2. Calculate the semantics of each title / indexed item (targets from the standard string based match with the query but before retrieval of the associated content) and calculate semantic matching or overlap of the query with each target. Then select the target(s) with high semantic matching. In this case new search method using semantic matching (instead of string matching) will be needed. Searching with semantic meaning will retrieve very appropriate and much less information.

3. Extract the most appropriate and related results from the search results in step #2.

Notes:
(a) Approach A and B can also be combined to get more appropriate results for some applications.
(b) Search algorithm of existing search engines may be modified to retrieve not all string matched content but retrieve contents only with high semantic match after doing a semantic match with the potential targets.

7.2 An Intelligent Summarization (IS) system

Semantics for multiple sentences and paragraphs can be calculated using SEBLA as it was done for calculating semantics of a sentence. However, some modifications are need for the following reasons:

1. Within a sentence, words are used in a constrained way using grammar. But between sentences there is no such grammar.

2. Usually, a group of sentences carry a theme within a context and there are ***relations*** between sentences.

Thus, to calculate the semantics between sentences, we will use word semantics as before BUT with some modifications. This is also true for a single long sentence segmented by “comma”, “semicolon”, “but”, “as” and the like. We also need to take account **for “discourse” i.e. coherence or co-reference to words in previous sentences**. There are some good existing solutions mainly for a small domain problem. But, in general Computational Discourse (CD) in natural language is an unsolved problem. However, with our SEBLA based scheme, the CD problem can be solved to a good extent for large domains.

In calculating semantics in a long sentence, the previous, next and other words can further influence / refine the semantics. For convenience, we have included this aspect in calculating semantics of multiple sentences. Fig. 4 shows the algorithm for summarization.

Apart from these two broader type applications, there are many specific applications of Big Data, e.g. bioinformatics, city planning, modeling climate change, deriving new economic theories, analysis reports, planning, classifying documents, and many more.

VIII. FUTURE WORKS

We plan to develop more enhanced Intelligent Agent (IA) and SEBLA for more complex Big Data applications, especially for Intelligent Search, better Q & A System, “Summarization” and “Converting data to information to knowledge to intelligence”. We also plan to work on using SEBLA to handle large structured data and develop Intelligent Database using semantics.

Our plan also includes developing various specific applications including analytics tools.

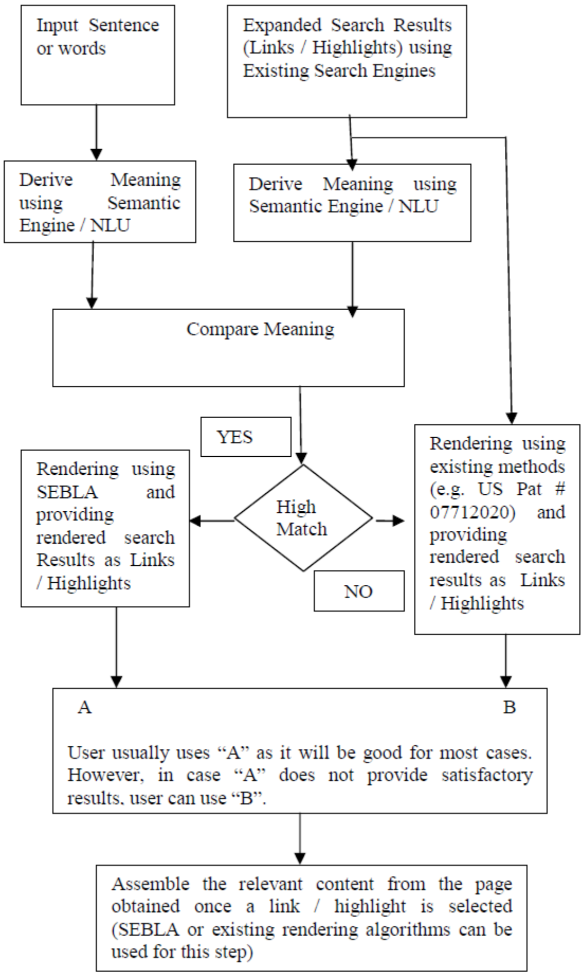

Fig. 3 Details of the block in Fig. 2 that shows Rendering And Assembling the Most Desired Content

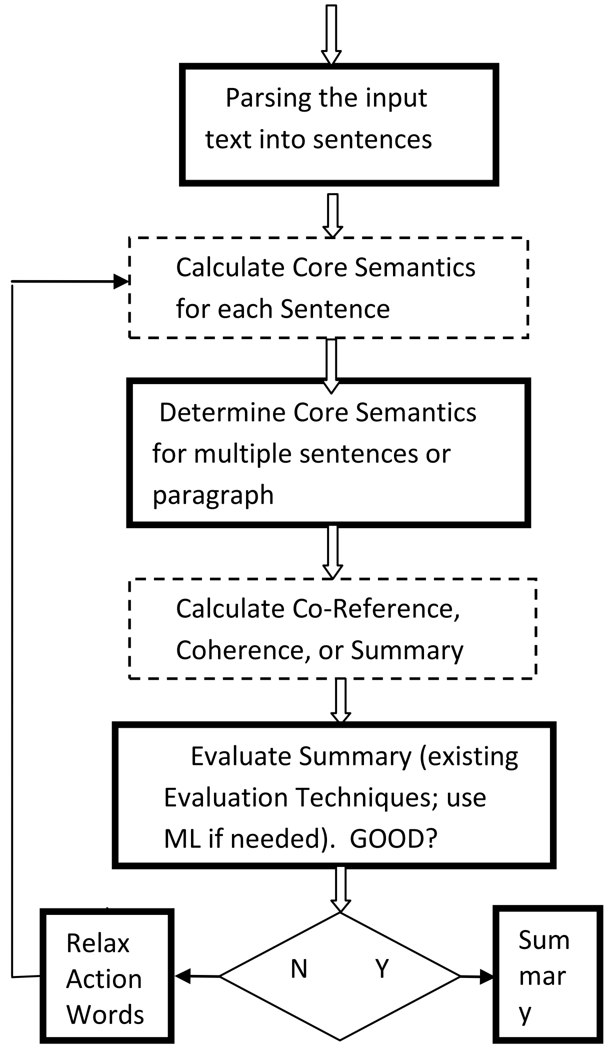

Fig. 4: Calculating Semantics for Multiple Sentences. If Summarization is calculated, the feedback loop is used for further refinement when needed.

IX. CONCLUSIONS

It is important to address the key Big Data problems in an effective and efficient way. We have discussed how to address Big Data problems using semantics and Natural Language Understanding (NLU). We have shown how Semantic Engine using Brain-Like Approach (SEBLA) and NLU can address key problems with big data including storage, search, transfer, sharing, analysis, processing, viewing, and deriving meaning / semantics.

We have focused mainly on the unstructured data. However, our approach is applicable for structured data as well as briefly explained.

We have emphasized on human Brain-Like and Brain-Inspired algorithms as humans can significantly compress the data by representing with a few words or sentences using the semantics of the information while keeping the core meaning or semantics. Humans use hierarchical multi-level compression of the sentences, paragraphs, pages using the semantics. SEBLA and NLU handle big data in a similar way. The main theme in SEBLA is to use each word as object with all important features, most importantly the semantics. The next main theme is to use the semantics of each word to develop the meaning of a sentence as we do in our natural language understanding as human. Similarly, the semantics of sentences are used to derive the semantics or meaning of a paragraph. The 3rd main theme is to use natural semantics as opposed to existing "mechanical semantics" of Predicate logic or Ontology or the like.

Solutions presented in this paper would provide much better solution for summarization, drawing inference or converting data to knowledge and to intelligence. This, accordingly, would help increase Data equity. Solutions presented are also good for various other applications including Intelligent Information Retrieval (IIR), Intelligent Search, Question and Answer System, and more accurate Language Translation.

REFERENCES

[1] C. Eaton et al, "Understanding Big Data: Analytics for enterprise class Hadoop and Streaming Data", http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF

[2] T. White, "*Hadoop: The Definitive Guide*", O'Reilly Media. p. 3. ISBN 978-1-4493-3877-0.

[3] J. Dean et al, "MapReduce: Simplified Data Processing on Large Clusters", OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[4] Wikipedia – "Big Data" - http://en.wikipedia.org/wiki/Big_data

[5] P. Ryan et al, "The Problem of Analyzing Unstructured Data", Grant Thoronton, 2009, http://www.grantthornton.ie/db/Attachments/Publications/Forensic_&_inve/Grant%20Thornton%20-The%20problem%20of%20analysing%20unstructured%20data.pdf

[6] E. Khan, " Intelligent Internet: Natural Language and Question & Answer based Interaction", Journal of Computers and Communications NAUN & UNIVERSITY PRESS), October 2013.

[7] E. Khan, "Natural Language based Hman Computer Interaction: a Necessity for Mobile Devices ", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) Dec. 2012.

[8] Khan, E., (2011): Natural Language Understanding Using Brain-Like Approach: Word Objects and Word Semantics Based Approaches help Sentence Level. A Patent Filed in US in 2011.

[9] E. Khan, "Internet for Everyone – Reshaping the Global Economy by Bridging the Digital Divide", Book- ISBN 978-1-4620-4251-7(Soft Cover), 978-1-4620-4250-0 (Hard Cover), Aug 2011.

[10] E. Khan & E. Aleisa, "e-Services using any Phone & User's Voice: Bridging Digital Divide & help Global Development", IEEE International Conference on Information Technology and e-Services, March 24-26, 2012, Tunisia.

[11] E. Khan, "Information for Everyone using any Phone –…", International Convention on Rehab. Engg. & Assistive Technology in Collaboration with ACM, July 2010, Shanghai, China.

[12] C. N. Hammack et al, "Automated Ontology Learning for a Semantic Web", Dept of CS, Uni. Of Nebraska, Feb 2002.

[13] E. Khan, "Addressing Big Data Problems using Semantics and Natural Language Understanding", 12th WSEAS International Conference on TELECOMMUNICATIONS And INFORMATICS (TELE-INFO '13) in Baltimore, MD, USA, September 17-19, 2013.

[14] [Wkiki-Predictive Analytics] http://en.wikipedia.org/wiki/Predictive_analytics

[15] E. Khan et al, "Internet Access by Talking and Listening Using any Phone: Help Global Development by Effectively Bridging the Digital and Language Divides", *International Journal of Information Studies*, Vol. 4, No 1, 2012.

[16] E. Khan," Method for Facilitating Internet Access with Vocal and Aural Navigation, Selection and Rendering of Internet Content", *US Pat #* 07712020, May 4, 2010.

[17] S. Pastore, "The Technology Evolution in Web Design, Development and Deployment Supporting Web Scientific Applications", *International Journal of Computers and Communication*, Issue 1, Volume 5, 2011.

[18] A. Rajaraman et al, "Mining Massive Dataset", Cambridge University Press, 2012, **,** ISBN-10: 1107015359 | ISBN-13: 978-1107015357.

[19] Computing Research Association, "Challenges and Opportunities with Big Data", A community White Paper developed by leading researchers in USA: www.cra.org/ccc/files/docs/init/**bigdatawhitepaper**.**pdf.**

[20] Cyber, "Data Equity: Unlocking the Value of Big Data", A Report for SAS**,** www.sas.com/offices/europe/uk/downloads/**data-equity**-cebr.pd