Addressing Bioinformatics Big Data Problems using Natural Language Processing: Help Advancing Scientific Discovery and Biomedical Research

EMDAD KHAN
College of Computer & Information Science
Imam University
Riyadh
SAUDI ARABIA
emdad@ccis.imamu.edu.sa

Abstract: - The amount of data in Bioinformatics (and amount of data in our world in general) has been exploding. For example, U.S. healthcare industry alone had generated 150 exabytes (2^18) of data by 2011. Using such large data sets - so called **big data** - has become a critical issue providing both **challenges and opportunities**. There are multiple problems with big data including storage, search, transfer, sharing, analysis, processing, viewing, deriving meaning / semantics, and drawing inference / converting data to knowledge. Hence, the need to solve these key problems related to Big Data in a practical and effective way is becoming very important.

Converting Big Data to "Knowledge" is becoming increasingly important to get real benefits from Big Data. It is claimed that U.S. healthcare industry alone can save $450 billion a year with the help of advanced analytics. In this paper, we propose Semantic Engine using Brain-Like Approach (**SEBLA**) and associated Natural Language Understanding (**NLU**) based approach to address the key problems of big data in bioinformatics and biology. Our approach resembles 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 preserving the core meaning. Thus, it very effectively converts data to knowledge and also compresses it; and hence addresses the key Big Data problems in an effective way. We describe how SEBLA and NLU can be used to handle both unstructured and structured data for addressing complex problems including **analytics**, understanding **biological systems/processes** (e.g. Gene Expression, Gene Function, and Protein Scaffolding) and **modeling biological systems/ processes**.

Key-Words: - Bioinformatics; Biology; Big Data; Unstructured Data; Natural Language Processing (NLP); Semantics; Intelligent Agent; Predictive Analysis; Business Intelligence; Biological Systems Modeling.

1 Introduction

The advent of DNA sequencing methods has greatly accelerated biological and medical research and discovery. The DNA sequencing cost has come down significantly along with the time to complete it. Knowledge of DNA sequences has become indispensable for basic biological research, and in numerous applied fields such as diagnostic, biotechnology, forensic biology, and biological systematics. The rapid speed of sequencing attained with modern DNA sequencing technology has been instrumental in the sequencing of complete DNA sequences, or genomes of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species [1].

What are the next key problems in biological systems / processes? There are 3 broad classes, namely, **analyzing & understanding biological systems, modeling biological systems / processes, and analytics.** Key problems under the **"**analyzing & understanding biological systems" are: understanding Gene Expression, Gene Function, Protein Scaffolding, Metabolism and the like. Modeling biological processes / systems is the key to address the issues under the first category. In fact, use of computational modeling is at the heart of systems biology. Although significant advancements have been made in modeling biological systems, it has long way to go. Today, there is no reliable and complete way to model a genetic network (e.g. Circadian clocks that provide endogenous cellular rhythms of approximately 24 hours that control many physiological processes), cells, organs, diseases (e.g. diabetes, cancer,..) and biological systems ([2], [3]). Solving the analytics problem in an automated way is the key as there are vast amount of literatures which is also growing very rapidly. Processing such literatures even at the initial stage of categorizing or grouping would provide a significant help. And, of course, summarization and drawing inferences in an automated or semi-automated way would be of great help in advancing the research.

Bioinformatics helps understanding of biological systems using computer science, especially to understand how information is **represented and transmitted** in biological systems. Bioinformatics is the key to help understand genomics, proteomics, biological processes, system biology, complex diseases (e.g. diabetes, cancer), drug discovery and more. ***Many aspects of computer science become handy including databases & database management, search engines, data visualization, NLU/NLP algorithms, machine learning, data mining, pattern matching, modeling and simulation.***

Due to the very large data size, the issues of Big Data come into play strongly in addressing most of the problems associated with biological systems. Big data in medical research is transforming research from hypothesis-driven to data-driven. Efficient analysis and interpretation of big medical data can open up new avenues to explore, new questions to ask, and new ways to answer, leading to better understanding of diseases and development of better and personalized diagnostics and therapeutics [4]. Thus, addressing big data problems in bioinformatics (and in biology) plays a critical role in turning data into meaningful biological applications and knowledge; thus help addressing all above mentioned three major categories of problems at hand, as well as help advancing the research.

In this paper, we discuss the use of Natural Language Processing (NLP), Natural Language Understanding (NLU) & associated semantics to address above mentioned Big Data based key problems. Our main focus is on the 1st and 3rd categories of problems i.e. **analyzing & understanding biological systems and doing automated analytics using NLU/NLP & Intelligent Agents.** Section 2 describes “NLP and Bioinformatics”. Section 3 describes “Solving Unstructured Big Data” and Section 4 describes **“**Solving Structured Big Data” in Bioinformatics / Biology. Section 5 discusses how NLP (along with semantics) is used as a key element to help understand Biological systems. Section 6 focuses on Future Works with special emphasis on modeling biological systems using NLP, and Section 7 provides conclusions.

2 Natural Language Processing (NLP) and Bioinformatics

To handle Big Data in bioinformatics, biomedical informatics and biology (and Big Data in general), we would need some automated method as it is not possible for human to manually try to process, understand and derive new inferences from such large amount of data. Big Data consists of unstructured (free text data) and structured data (e.g. data in a database). Unstructured data dominates the data world. It is estimated that over 80% data in computers and Internet are unstructured [6]. In case of bioinformatics, the structured data is also very large - e.g. data in MEDILINE and GenBank. Computers are very good in processing structured data. 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 [7]. Semantics is also very key to improve the usage of structured data – in finding relations, extracting new information and connecting / using structured data with unstructured data [8]. Thus, Natural Language Processing (NLP) and associated semantics become very useful in addressing Big data problems in bioinformatics and biology. In fact, use of NLP in biology has been increasing rapidly. A very good description of how NLP is used for Information Management in biology and bioinformatics is provided in [9]. In [10], Semantic MEDLINE integrates information retrieval, advanced natural language processing, automatic summarization, and visualization into a single Web portal. Semantic MEDLINE can make an impact on biomedicine by supporting scientific discovery and the timely translation of insights from basic research into advances in clinical practice and patient care.

It is important to note that although existing NLP approaches have made good progress and simplified the automation process somewhat, 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. In case of bioinformatics, existing methods mainly do information management (information retrieval and information extraction). The capabilities to reliably finding relationships between genes / proteins, generating specific predictions that pertain to gene function, predicting essential genes, and finding correct interactions are limited. E.g. co-occurrence of gene and protein names in abstracts implies a biological relationship. But in many cases co-occurrences are not indicative of interaction. Negation is one trivial reason (e.g. A was found not to interact with B [9]). Use of controlled vocabulary in today's ontology is another key limitation. E.g. an author may refer to "type II diabetes mellitus" but an ontology concept may consider this as "diabetes, type II, mellitus" which usually cause major difficulty for a software used to search texts (not a big issue for humans though).

The key point is that we would need to use better semantics and NLU capabilities in dealing with both unstructured and structured data to more reliably and efficiently address such issues. In [8], we proposed to use Semantic Engine using Brain-Like Approach (SEBLA) to convert data to knowledge and also to compress it; thus addressing the Big Data problems in an effective way. SEBLA provides "Natural Semantics" i.e. semantics similar to what humans use (see Section 3 for more details). Due to the natural semantics capability of SEBLA, more complex cases can be addressed e.g. understanding **biological problems** (e.g. Gene Expression, Gene Function, and Protein Scaffolding) and help modeling biological processes / systems (Sections 5 and 6).

Below is a brief description of how NLP with better semantic capability can address various problems including Business Analytics (BI), Information Management, Understanding Biological Systems and Modeling Biological Systems.

2.1 Analytics

Analytics, in general, is a process to analyze large data, discover meaningful patterns and then draw some inferences as well as do summarization. It is usually done for business intelligence (BI). But, the same concept can be applied in biology and bioinformatics to do Research Intelligence (RI) i.e. similar to Business Intelligence. In addition to using NLP for information management to retrieve and extract important information, we also need to do summarization and draw some good inferences from large biological data. This also includes filling some structured data tables (e.g. tables in a database) using relevant data from vast amount of text data. Understanding key research issues, research trends etc are important to advance the research more effectively. The same can be applied to medical, biomedical, biological and bioinformatics business intelligence.

2.2 Information / Knowledge Retrieval, Extraction and Integration from various sources

There are various sources for genomics and proteomics information. In general, such sources use different styles, formats even though most use common ontology like in Genome Ontology (GO). Correctly retrieving, extracting and integrating information from such sources is the key to better analyze, understand and derive new information. This mainly belongs to information management (i.e. information retrieval, extraction and associated alignment). NLP has made great progress in this area, especially, exploring and managing biomedical literature [9]. The flood of sequence information produce by the rapid advances in genomics and proteomics is a key driver in bringing the use of NLP to bioinformatics. The fact that so many texts and sequences are available now electronically, it is clear that NLP become an obvious choice of extracting key information from such vast sources.

From information management standpoint, NLP has 3 aspects: information retrieval, information extraction and semantics. Information retrieval refers to the recovery of documents from databases related to user's query (e.g. use of PubMed to find documents about a topic). Search from the Internet and databases can be grouped under Information Retrieval. The goal is to find the most related information to the query. This is probably the most common use of NLP today. Existing information retrieval methods are mainly based on string matching.

Information extraction is the process of retrieving some meaning from a text – for example, finding protein-protein interaction from MADELINE. String based extraction is not useful to extract meaning, hence technologies like ontologies, parsing (syntactic and semantic) and regular expressions are needed.

Semantics (i.e. the meaning of words and sentences) is the critical element for information extraction. It is also an important element for much better information retrieval. Semantic search can provide much more relevant and much concise search results. However, semantics based on exiting methods (e.g. ontologies) may not produce key information for many cases as just structural relationships between words do not convey the core meaning in many cases (refer to Sections 3 and 5 for more details). As mentioned, natural semantics based semantic engine SEBLA can improve information extraction and retrieval in a major way.

2.3 **Understanding Biological system**

Information retrieval and extraction using vast sources of data is very important to automatically process Big Data, and help understanding of biological systems by the researchers mainly from a **higher level**. However, we believe, NLP and NLU using semantics can be used to better understand the biological systems and processes at deeper levels – e.g. to understand Gene Function, Gene Expression, Genetic Messages, Protein Scaffolding, and Metabolism. This is because biological systems use **biological alphabets** in Genes and Proteins. Thus, finding special sequences of such alphabets and words, their relations and drawing some good inferences are keys to understand biological systems. And these are closely related to NLP & NLU.

2.4 Developing Semantics in Biological systems

There is a big caveat for the concept described above in Section 2.3. Biological words (e.g. 3 letter words [codon] in a DNA sequence) are not like our natural language words for which we know the complete meaning. Only biological systems know the real meaning and vocabulary of such words. However, we believe, we should be able to use SEBLA's natural semantics approach to develop semantics of biological words and then apply it to understand biological systems and processes. It is important to note that only about 2% of total bases in a gene are used to code proteins. We do not know what exactly the remaining 98% of the gene are doing. As per [16], only about 1% of the three billion letters directly codes for proteins - of the rest, about 25% make up genes and their regulatory elements. The function of the remaining letters is still unclear. Some of it may be redundant information left over from our evolutionary past. Existing methods usually involve comparing new sequences with existing one, discovering structure and function by homology (the existence of shared ancestry between a pair of structures, or genes, in different species) rather than through a true understanding of the biological principles underlying structure and function. We believe such problems can be addressed using NLP/NLU principles after developing the semantics in biological systems. If successful, this would also help better understand the evolution process.

2.5 Modeling Biological Systems

Modeling biological processes / systems is the key to better understand such processes / systems, do deeper analyses, discover new information and draw valuable inferences. This will significantly help advance the research, drug discovery, personalized medicine and more. Semantics of NLU can also play a major role in modeling biological systems as briefly described in Section 6.

3 Semantics and NLU to Address Unstructured Big Data Problems

The key problems associated with unstructured data 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 ([11], [12]) using Brain-Like algorithms (SEBLA). Then we show how SEBLA can handle Big Data in bioinformatics.

3.1 Semantic Engine Using Brain-Like Approach (SEBLA)

While NLP / NLU are widely used, their success so far have been mainly in a small domain. For large domain and from semantic standpoint, 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 [14], 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, 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.

Although above example shows the issue of existing semantics using a Question & Answer type system, the same applies for almost all cases including information retrieval, search and information extraction.

3.2 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 from biological context can be broadly classified as:

a. Information Retrieval (IR) / Search
b. Information Extraction
c. Question & Answer
d. Summarization
e. Converting data to information to knowledge to intelligence

[*Note: as mentioned above, semantics and NLU/NLP are also important to understand and model biological systems – these aspects are described in Sections 5 and 6*]

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.

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. 1.

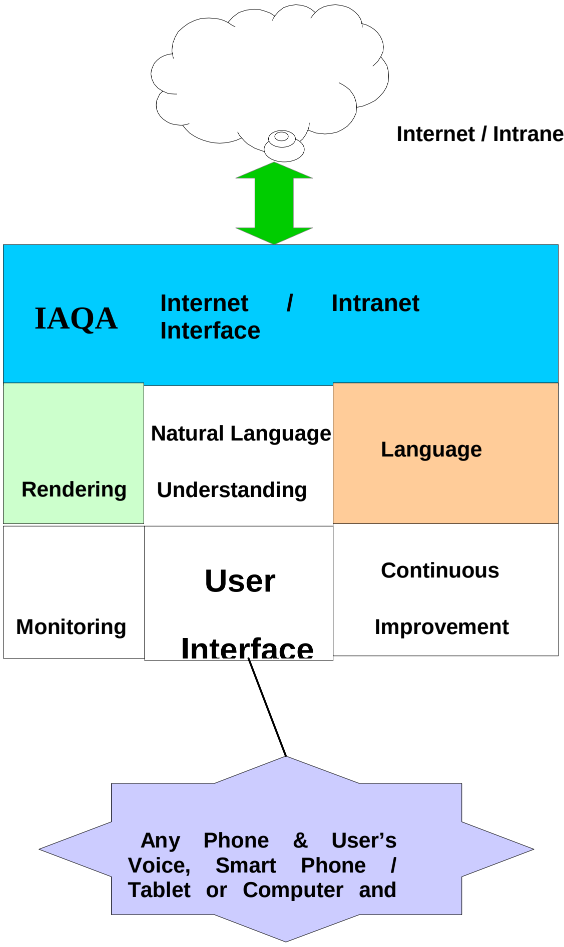
Internet / Intranet
IAQA
Internet / Intranet Interface
Rendering
Natural Language Understanding
Language
Monitoring
User Interface
Continuous Improvement
Any Phone & User's Voice, Smart Phone / Tablet or Computer and

Fig. 1 IAQA: Intelligent Agent for a Question & Answer (Q & A) System.

The term "**rendering**" ([12] [13]) 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.

4. Semantics and NLU to Address 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 already very large for bioinformatics / biology and also growing very fast. Conventional algorithms are not sufficient for 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 are critical to process them effectively and efficiently.

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 (e.g. MEDLINE, GenBank). 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.

5. NLP as a key element to Help Understand Biological systems

The use of NLP to help understand biological systems and processes is already described in Section 2.3. There are two broad categories:

a. Use Big Data to understand at a higher level. This is basically automatic use of Big Data inferences by the researchers. Due the nature of Big Data and the information that can be inferred, this can be a great contributor to researchers to better understand biological systems.

b. Applying NLP / NLU concept to biological language consisting of biological alphabets in genes (A, T, C, G), proteins (ALA, ARG, ASP,….) and words.

Use of Big Data to better understand biological systems at a higher level is explained in Section 2.3. Application of NLP / NLU concept to help understand biological systems / processes at deeper levels is also mentioned in Section 2.3. However, it needs more explanation. At the first level, basics of NLP (e.g. Regular Expressions, String processing, String search, and pattern analysis) can be used to retrieve new information from large data set. This will help in finding similar genes, finding closest neighbor of a new gene, what specific patterns in gene sequence results 3-D shape of proteins and the like.

The next level is determining the real meaning of the genetic words and sentences (sequence of words) using the semantics. This will help us to really understand the genetic messages, how biological subsystems and systems work. It will help us to understand the general complete biological process (equation (1)) i.e.

Genetic Information -> Molecular Structure -> Biochemical Function -> Biological Behavior ……. (1)

It will also possibly help to understand the major part of the gene (about 74%, [16]) that is not understood yet. However, as discussed in Section 2.4, we would need to develop the semantics first which may be a daunting task. But it is surely worth pursuing.

6. Future Works

We plan to develop a complete BI (business Intelligence) / RI (Research Intelligence System) using SEBLA based NLU. We also plan to develop semantics of biological basic words and sentences (sequence of words) by using the knowledge how biological systems work (as much as we know today) and associated Big Data. We will then apply such findings and NLU

(a) to better understand how the biological systems and processes works via the semantics that will be developed.

(b) we will also try to model biological systems using the understanding developed via semantics.

(a) and (b) will help each other to further refine, and better understand as well as to better model biological systems and processes.

7. Conclusions

We have presented Semantic Engine using Brain-Like Approach (**SEBLA**) and associated Natural Language Understanding (**NLU**) based approach to address the key problems of Big Data in bioinformatics and biology. We have used human Brain-Like and Brain-Inspired algorithms as humans can significantly compress the data, preserve core meaning, extract latent information, and convert information to knowledge and intelligence. Thus, Brain-Like approach very effectively converts data to knowledge and also compresses it; and hence addresses the key Big Data problems in an effective way. We presented how SEBLA and NLU are used to handle both unstructured and structured data for addressing complex problems including **analytics**, understanding **biological systems/processes** (e.g. Gene Expression, Gene Function, and Protein Scaffolding) and **modeling biological systems/ processes**.

We have emphasized that use of NLP / NLU along with associated semantics is the key to understand biological systems and processes as biological systems use biological alphabets, words and sentences (sequence of words) similar to our natural language. The key difference is that we know the semantics and meaning of words and sentences in our natural language, but we do not know such semantics for the words and sentences used in the biological systems and processes.

We believe we can develop the semantics of biological basic words and sentences (sequence of words) by using the knowledge how biological systems work (as much as we know today) and associated Big Data. We can then apply such findings and NLU to better understand how the biological systems and processes work via the semantics. We believe this will also help better modeling of biological systems. Such efforts, if successful, will enable us to not only understand how biological systems / processes really work but also to understand the evolution and other hidden functions / processes as the functions of about 74% of bases in a gene would be understood.

References:

[1] Wikipedia – "DNA Sequencing" – http://en.wikipedia.org/wiki/DNA_sequencing.

[2] R. Schwartz, "Biological Modeling and Simulation", ISBN 978-0-262-19584-3, MIT Press, 2008.

[3] Z. Azallasi et al, "System Modeling in Cellular Biology", ISBN 978-0-262-19584-5, MIT Press, 2008.

[4] Big Data Initiative by U.S. President Obama - http://www.whitehouse.gov/blog/2013/04/23/big-data-big-deal-biomedical-research.

[5] 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

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

[7] 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

[8] 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.

[9] M. Yandell et al, "Genomics and Natural Language Processing", Nature Reviews (Genetics), Vol. 3, Aug 2002.

[10] H. Kilicoglu et al, "Semantic MEDLINE: A Web Application for Managing the Results of PubMed Searches", Journal of Information Services and Use, IOS Press, Vol. 31, #1-2, Aug 11, 2011.

[11] E. Khan, "Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach**",** submitted to Journal– acceptance expected by Dec. 2013 Jan 2014.

[12] E. Khan, " Intelligent Internet: Natural Language and Question & Answer based Interaction", INTERNATIONAL JOURNAL of COMPUTERS AND COMMUNICATIONS, (NAUN & UNIVERSITY PRESS) Oct. 2013.

[13] Internet for Everyone - Reshaping the Global Economy by Bridging the Digital Divide", Book - ISBN 978-1-4620-4251-7 (SC ISBN)978-1-4620-4250-0 (HC ISBN), Aug 2011.

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

[15] D. Brutlag et al, "Understanding Human Genome", Scientific American: Introduction to Molecular Medicine, 1994.

[16] "DNA Molecule: How Much DNA Codes for Protein?" http://www.dnalc.org/resources/3d/09-how-much-dna-codes-for-protein.html, April2, 2010.