BIG DATA ANALYSIS USING SP THEORY OF INTELLIGENCE

**Mala S1, Varsha C G2,Chaithra B S3 ,Gaurav4**

ABSTRACT

SP Theory of intelligence and its realization in the SP machine may with the advantage be applied to the management and analysis of big data.The SP system helps to overcome the problem of variety,volume and velocity in big data;it has potential as a universal framework for the presentataion and processing of diverse kinds of knowledge,helping to reduce the diversity of formalisms and formats for knowledge,and the different ways in which they are processed.It lends itself to the analysis of straming data,helping to overcome the problem of velocity in big data.Central workings of the system are lossless compression of information making big data smaller and reducing problems of storage and management.There is potential for substantial economics in the transmission of data,for big cuts in the energy in computing,for faster processing,and for smaller and lighter computers.The SP system provides a way to handle the problem of veracity in big data,with potential to assist in the management of errors and uncertainties in data.It lends itself to the visualization of knowledge structures and inferential processes.

1. INTRODUCTION

Big data—the large volumes of data that are now produced many fields—can present problems in storage, transmission, and processing, but their analysis may yield useful information and useful insights.

Big data is defined as large amount of data which requires new technologies and architecture to make possible to extract value from it by capturing and analysis process. New sources of big data include location specific data arising from traffic management and from the tracking of personal devices such as smart phones. Big data has emerged because we are living in a society which makes increasing use of data intensive technologies. Due to much large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Since big data is recent upcoming technology in which market can bring large benefits to the business organizations.

Big data is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications. The difficulties can be related to data capture, storage, search, sharing, transfer, analysis and visualization. Big data due to its various properties like volume, velocity, variety, variability and complexity its forward in many challenges. The Various challenges faced in large data management include Scalability, unstructured data, accessibility, real time analysis, fault tolerance and many more. In addition to that variation in the amount of data stored in different areas, the type of data generated and stored such as images, video, audio or text/numeric information.

* 1. ISSUES IN BIG DATA

The issues in big data are related to the characteristics

**Data volume:** Due to increase in the volume of data, the value of different data records will decrease in type, age richness and quantity. The social networking site existing are themselves producing the data in order of terabytes everyday and is amount of data is definitely difficult to handle by using the existing traditional system.

**Data Velocity:** Our Traditional system is not capable enough on performing the analytics on data which is constantly in action. E-commerce has rapidly increase the speed and richness of data which is used for different business transaction such as website usage. Data velocity issues increases to manage the bandwidth limit.

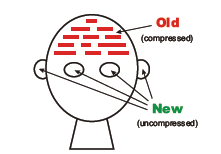
**Data Variety:** All this data are different type which consist of raw data, structured, unstructured and semi structured data which is difficult to handle by using the existing traditional analytic system from an analytic perspective it’s probably the biggest obstacle to effectively using large volume of data. Incomparable data format, non aligned data structure and inconsistent data, semantics represent significant challenges that can lead to analytic sprawl.

**Storage and Transport Issues:** The most recent data explosion mainly due to social media. Moreover the data is created by professionals, such as scientist, journalist, writers and from mobile device to super computers, there is no sufficient storage devices to store this large data. Currently disk technology limits are about 4 terabytes per disk. Even if 1 Exabyte’s of data could be processed in a single system, it is not able to attach directly.

**Data Management & Processing Issues:** The most difficult problem to address with big data are resolving the issues of access, utlilization, updating, goveranance and reference are proven. The sources of data are varied by size, format. The effective processing of exabytes of data will require extensive parallel processing and analytics algorithms in order to provide timely and actionable information.

1.2 SP THEORY OF INTELLIGENCE & SP MACHINE

The SP Theory of Intelligence, which has been under development since about 1987 aims to Simplify and Integrate concepts across Artificial Intelligence, Mainstream Computing and Human Perception and Cognition, with Information Compression as a unifying theme. The name “SP” is short for Simplicity and Power, because compression of any given body of information, I, may be seen as a process of reducing informational “redundancy” in I and thus increasing its “simplicity”, whilst retaining as much as possible of its non-redundant expressive “power”. Likewise with Occam’s Razor



The theory is conceived as an abstract brain-like system that, in an ‘input’ perspective, may receive New information via its senses, and compress some or all of it to create Old information, as illustrated schematically in Fig. In the theory, information compression is the mechanism both for the learning and organisation of knowledge and for pattern recognition, reasoning, problem solving, and more. In the SP system, all kinds of knowledge are represented with patterns: arrays of atomic symbols in one or two dimensions. At the heart of the system are processes for compressing information by finding good full and partial matches between patterns and merging or ‘unifying’ parts that are the same.

The close association between information compression and concepts of prediction and probability means that the SP system is intrinsically probabilistic. Each SP pattern has an associated frequency of occurrence, and for each multiple alignment, the system may calculate associated probabilities.Although the SP system is fundamentally probabilistic, it can, if required, be constrained to operate in the clockwork style of a conventional computer, delivering all-or-nothing results

An important idea in the SP programme is the DONSVIC principle [3, Section 5.2]: the conjecture, supported by evidence, that information compression, properly applied, is the key to the discovery of ‘natural’ structures, meaning the kinds of things that people naturally recognise, such as words, objects, and classes of objects. Evidence to date suggests that the SP system does indeed conform to that principle.

The SP theory is realised in a computer model, SP70, which may be regarded as a first version of the SP machine. It is envisaged that the SP computer model will provide the basis for the development of a high-parallel, open-source version of the SP machine.The theory has things to say about several aspects of computing and cognition, including unsupervised learning, concepts of computing, aspects of mathematics and logic, the representation of knowledge, natural language processing, pattern recognition, several kinds of reasoning, information storage and retrieval, planning and problem solving, and aspects of neuroscience and of human perception and cognition.

In the SP theory, information compression is the mechanism both for the learning and organization of knowledge and for pattern recognition, reasoning, problem solving.

The existing and expected benefits of the SP theory and some of its potential applications.

• Conceptual simplicity combined with descriptive and explanatory power.

• Simplification of computing systems, including software.

• Deeper insights and better solutions in several areas of application.

• Seamless integration of structures and functions within and between different areas of

application.

In broad terms, the SP theory has three main elements:

• All kinds of knowledge are represented with patterns: arrays of atomic symbols in one or two dimensions.

• At the heart of the system is compression of information via the matching and unification (merging) of patterns, and the building of multiple alignments.

• The system learns by compressing “New” patterns to create “Old” patterns

An important idea in the SP programme is the DONSVIC principle the conjecture, supported by evidence, that information compression, properly applied, is the key to the discovery of ‘natural’ structures, meaning the kinds of things that people naturally recognize, such as words, objects, and classes of objects. Evidence to date suggests that the SP system does indeed conform to that principle. The SP theory is realized in a computer model, SP70, which may be regarded as a first version of the SP machine. It is envisaged that the SP computer model will provide the basis for the development of a high-parallel, open-source version of the SP machine.

1. PROPOSED SYSTEM

In a Proposed SP System, it is designed to simplify and integrate concepts across artificial intelligence, mainstream computing, and human perception and cognition, has potential in the management and analysis of big data. The SP system has potential as a universal framework for the representation and processing of diverse kinds of knowledge (UFK), helping to reduce the problem of variety in big data the great diversity of formalisms and formats for knowledge, and how they are processed. The system may discover ‘natural’ structures in big data, and it has strengths in the interpretation of data, including such things as pattern recognition, natural language processing, several kinds of reasoning, and more. In the Broad potential benefits of the SP system, as applied to big data, are in these areas:

**Overcoming the problem of variety in big data.**

Harmonizing diverse kinds of knowledge, diverse formats for knowledge, and their diverse

modes of processing, via a universal framework for the representation and processing of knowledge.

**Interpretation of data**.

The SP system has strengths in areas such as pattern recognition, information retrieval, parsing and production of natural language, translation from one representation to another, several kinds of reasoning, planning and problem solving.

**Velocity- Analysis of Streaming Data.**

The SP system lends itself to an incremental style, assimilating information as it is received,

much as people do.

**Volume - Making Big Data Smaller.**

Reducing the size of big data via lossless compression can yield direct benefits in the storage, management, and transmission of data, and indirect benefits in several of the other areas.

* 1. INTERPRETATION OF DATA

By contrast with unsupervised learning, which compresses a body of information (I) to create G and E, the concept of interpretation in this article means processing I in conjunction with a pre-established grammar (G) to create a relatively compact encoding (E) .

Depending on the nature of I and G, the process of interpretation may be seen to achieve such things as pattern recognition, information retrieval, parsing or production of natural language, translation from one representation to another, several kinds of reasoning, planning, and problem solving.

Some of these were touched on briefly in Section..

**Pattern recognition**: With the SP system, pattern recognition may be achieved: at multiple levels of abstraction; with “family resemblance” or polythetic categories; in the face of errors of omission, commission or substitution in data; with the calculation of a probability for any given identification, classification or associated inference; with sensitivity to context in recognition; and with the seamless integration of pattern recognition with other aspects of intelligence—reasoning, learning, problem solving, and so on . As previously mentioned, the system may be applied in computer vision and in medical diagnosis , viewed as pattern recognition. \_

**Information retrieval** : The SP system lends itself to information retrieval in the manner of query-by-example and with the provision of SP patterns representing relevant rules, there is potential to create the facilities of a query language like SQL

**Parsing and production of natural language.**As can be seen in Fig, the creation of a multiple alignment in the SP system may achieve the effect of parsing a sentence in a natural language . It may also function in the production of sentences

**Translation from one representation to another.** There is potential with the SP system for the integration of syntax and semantics in both the understanding and production of natural language , with corresponding potential for the translation of any one language into an SP-style interlingua and further translation into any other natural language [5, Section 6.2.1]. Probably less challenging, as mentioned earlier, would be the translation of artificial formalisms and formats JPEG, MP3, and so on—into an SP-style representation.

**Several kinds of reasoning**. The SP system can perform several kinds of reasoning, including one- ‘deductive’ reasoning, abductive reasoning, reasoning with probabilistic networks and trees, reasoning with ‘rules’, nonmonotonic reasoning, Bayesian reasoning and “explaining away”, causal diagnosis, and reasoning that is not supported by evidence

**Planning**. With SP patterns representing direct flights between cities, the SP system can normally work out one or more routes between any two cities that are not connected directly, if such a route exists.

**Problem solving**. The system can also solve textual versions of geometric analogy problems, like those found in puzzle books and IQ tests

* 1. VELOCITY: ANALYSIS OF STREAMING DATA

“Most of today’s computing tasks involve data that have been gathered and stored in databases. The data make a stationary target. But, increasingly, vitally important insights can be gained from analyzing information that’s on the move. ... This approach is called streams analytics. Rather than placing data in a database first, the computer analyses it as it comes from a variety of sources, continually refining its understanding of the data as conditions change.This is the way humans process information.”

Although, in its unsupervised learning, the SP system may process information in batches, it lends itself most naturally to an incremental style. In the spirit of the quotation above, the SP system is designed to assimilate New information to a steadily-growing body of relatively-compressed Old information..Likewise, in interpretive processes such as pattern recognition processing of natural language, and reasoning, the SP system may be applied to streams of data as well as the processing of data in batches.

2.3 VOLUME: MAKING BIG DATA SMALLER

“Very-large-scale data sets introduce many data management challenges”.“In addition to reducing computation time, proper data representations can also reduce the amount of required storage (which translates into reduced communication if the data are transmitted over a network).” .Because information compression is central in how the SP system works, it has potential to reduce problems of volume in big data by making it smaller. Although comparative studies have not yet been attempted, the SP system has potential to achieve relatively high levels of lossless compression for two main reasons: it is designed so that, if required, it can perform a relatively thorough search for redundancies in data; and there is potential to tap into discontinuous dependencies in data, an aspect of redundancy that appears to be outside the scope of other systems for compression of information .

In brief, information compression in the SP system can yield direct benefits in the storage, management, and transmission of data, and indirect benefits as described elsewhere in this article: unsupervised learning , processes of interpretation such as pattern recognition and reasoning , economies in the transmission of data via the separation of grammar and encoding, gains in computational efficiency , and assistance in the management of errors and uncertainties in data .