**M.S.RAMAIAH INSTITUTE OF TECHNOLOGY**

**(Autonomous Institute, Affiliated to VTU)**

[www.msrit.edu](http://www.msrit.edu)

**BANGALORE-560054**



**Big Data Analysis Using SP Theory Of Intelligence**

**Submitted by**

Chaithra B S 1MS13CS403

Mala S 1MS13CS411

Varsha C G 1MS13CS422

Gaurav 1MS12CS141

**Guided by**

Soumyarani C N

Professor

Department of Computer Science and Engineering

**CONTENTS**

1. Abstract. - 3
2. Literature survey. - 4
3. Background study on the project. - 6
4. Problem statement and objectives. - 8
5. Methodology. - 11
6. Tools to be used. - 12
7. Application of the project. - 13
8. Expected outcomes. - 15
9. Conclusion. - 16

10. References. - 17

**ABSTRACT**

Big data includes large quantity of data that results drawbacks both in accessing and

Managing the data. These drawbacks can be overcome by introducing the SP theory of

Intelligence. The term ‘SP’ indicates ‘Simplicity’ and ‘Power’. The central concept used in the theory is lossless information compression. This helps in making the big data small, thereby provides benefits both in accessing and management. The purpose of this project is to overcome the problems in big data using the SP Theory of Intelligence. In order to achieve this goal, big data is subjected to compression techniques. Compression of information is achieved by pattern matching. Using such a system leads to the improvement in the processing of big data. The SP Theory provides pattern recognition, information storage, retrieval and information compression. Although this theory leads in faster information retrieval, the integrity of the original information is maintained. Future

work has to be done in this area to work with patterns in two dimensions. It has strengths in the unsupervised learning or discovery of structure in data, in pattern recognition, in the parsing and production of natural language. It lends itself to the analysis of streaming data, helping to overcome the problem of velocity in big data.**LITERATURE SURVEY**

**“SP theory of intelligence: An Overview”**

The theory aims to simplify and integrate concepts across artificial intelligence, mainstream computing and human perception and cognition, with information compression as a unifying theme. It is conceived as a brain-like system that receives 'New' information and stores some or all of it in compressed form as 'Old' information. It is realized in the form of a computer model -- a first version of the SP machine. The concept of "multiple alignment" is a powerful central idea. Using heuristic techniques, the system builds multiple alignments that are 'good' in terms of information compression. For each multiple alignment, probabilities may be calculated. These provide the basis for calculating the probabilities of inferences. The system learns new structures from partial matches between patterns. The system searches for sets of structures that are 'good' in terms of information compression. These are normally ones that people judge to be 'natural', in accordance with the 'DONSVIC' principle -- the discovery of natural structures via information compression. The SP theory may be applied in several areas including 'computing', aspects of mathematics and logic, representation of knowledge, natural language processing, pattern recognition, several kinds of reasoning, information storage and retrieval, planning and problem solving, information compression, neuroscience, and human perception. Examples include the parsing and production of language including discontinuous dependencies in syntax, pattern recognition at multiple levels of abstraction and its integration with part-whole relations, reasoning with default values, reasoning in Bayesian networks including 'explaining away’, and the solving of a geometric analogy problem.

**“Information compression, intelligence, computing, and mathematics”**

Paper presents evidence for the idea that much of artiﬁcial intelligence, human perception and cognition, mainstream computing, and mathematics, may be understood as compression of information via the matching and uniﬁcation of patterns. This is the basis for the SP theory of intelligence, outlined in the paper and fully described elsewhere. Relevant evidence may be seen: in empirical support for the SP theory; in some advantages of information compression (IC) in terms of biology and engineering; in our use of shorthand’s and ordinary words in language; in how we merge successive views of any one thing; in visual recognition; in binocular vision; in visual adaptation; in how we learn lexical and grammatical structures in language; and in perceptual constancies. IC via the matching and uniﬁcation of patterns may be seen in both computing and mathematics: in IC via equations; in the matching and uniﬁcation of names; in the reduction or removal of redundancy from unary numbers; in the workings of Post’s Canonical System and the transition function in the Universal Turing Machine; in the way computers retrieve information from memory; in systems like Prolog; and in the query-by-example technique for information retrieval. The chunking-with-codes technique for IC may be seen in the use of named functions to avoid repetition of computer code. The schema-plus-correction technique may be seen in functions with parameters and in the use of classes in object-oriented programming. And the run-length coding technique may be seen in multiplication, in division, and in several other devices in mathematics and computing. The SP theory resolves the apparent paradox of “decompression by compression”. And computing and cognition as IC is compatible with the uses of redundancy in such things as backup copies to safeguard data and understanding speech in a noisy environment.

**BACKGROUND STUDY ON THE PROJECT**

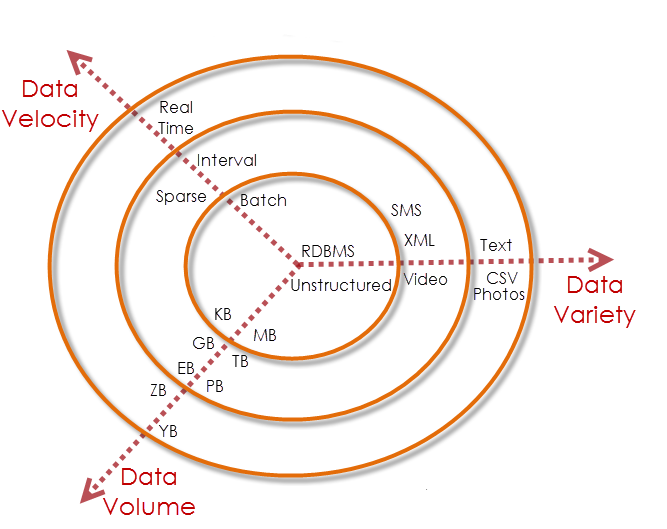
Big data the large volumes of data that are now produced in many fields can present problems in storage, transmission, and processing, but their analysis may yield useful information and useful insights.

**The 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.
* *Learning and discovery:* The unsupervised learning or discovery or discovery of ‘natural’ structures in data.
* *Interpretation of data:* The SP system has strengths inareas 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 systemlends itself to an incremental style, assimilating information as it is received, much as people do.
* *Volume: making big data smaller*. Reducing the sizeof 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 discussed in this article.
* *Additional economies in the transmission of data*. Thereis potential for additional economies in the transmission of data, potentially very substantial, by judicious separation of `encoding' and `grammar'.
* *Energy, speed, and bulk*. There is potential for big cutsin the use of energy in computing, for greater speed of processing with a given computational resource, and for corresponding reductions in the size and weight of computers.
* *Veracity:* *managing errors and uncertainties in data.* The SP system can identify possible errors or uncertainties in data, suggest possible corrections or interpolations, and calculate associated probabilities.
* *Visualization:* Knowledge structures created by the system, and inferential processes in the system, are al transparent and open to inspection. They lend themselves to display with static and moving images

**PROBLEM STATEMENT AND OBJECTIVES**

Simplicity and Power theory of intelligence to overcome the variety of problems in big data platform.

The 3 **V**s that define *Big Data* are **V**ariety, **V**elocity and **V**olume.

1. ***Volume*: Making Big Data Smaller**

* “Very-large-scale data sets introduce many data management challenges.”
* Information compression.
* Direct benefits in storage, management and transmission.
* Indirect benefits
* efficiency in computation and the use of energy
* unsupervised learning
* additional economies in transmission and the use of energy
* assistance in the management of errors and uncertainties in data
* Processes of interpretation.

**Energy, Speed and Bulk**

* In the SP theory, a process of searching for matching patterns is central in all kinds of ‘processing’ or ‘computing’.
* This means that anything that increases the efficiency of searching will increase computational efficiency and, probably, cut the use of energy:
* Reducing the volume of big data.
* Cutting out some searching.

**Efficiency via Reduction in Volume**

* Information compression is central in how the SP system works:
  + Reducing the size of big data.
  + Reducing the size of search terms.
* Both these things can increase the efficiency of searching, meaning gains in computational efficiency and cuts in the use of energy.

1. **Variety of Big Data**

* Diverse kinds of data: the world’s many languages, spoken or written; static and moving images; music as sound and music in its written form; numbers and mathematical notations; tables; charts; graphs; networks; trees; grammars; computer programs; and more.
* There are often several different computer formats for each kind of data. With images, for example: JPEG, TIFF, WMF, BMP, GIF, EPS, PDF, PNG, PBM, and more.
* Adding to the complexity is that each kind of data and each format normally requires its own special mode of processing.
* THIS IS A MESS! It needs cleaning up.
* Although some kinds of diversity are useful, there is a case for developing **a *universal framework for the representation and processing of diverse kinds of knowledge (UFK****).*

**Universal Framework for the Representation and Processing of Knowledge (UFK)**

* Potential benefits of a UFK in:
* Learning structure in data.
* Interpretation of data.
* Data fusion.
* Understanding and translation of natural languages.
* The semantic web and internet of things.
* Long-term preservation of data.
* Seamless integration in the representation and processing of diverse kinds of knowledge.
* Most concepts are an amalgam of diverse kinds of knowledge (which implies some uniformity in the representation and processing of diverse kinds of knowledge).
* The SP system is a good candidate for the role of UFK because of its versatility in the representation and processing of diverse kinds of knowledge.

**iii.**  **Velocity: Analysis of Streaming Data**

* In the context of big data, “velocity” means the analysis of streaming data as it is received.
* “This is the way human’s process information.”
* This style of analysis is at the heart of how the SP system has been designed.
* Unsupervised learning.

**METHODOLOGY**

* SP theory of intelligence.
* Accessing cloud.
* Encryption and decryption of data.

**TOOLS TO BE USED**

**HARDWARE REQUIREMENTS**

Processor : Pentium IV 2.4 GHz

Hard Disk : 250 GB

RAM : 1 GB

**SOFTWARE REQUIREMENTS**

Operating system : Microsoft windows

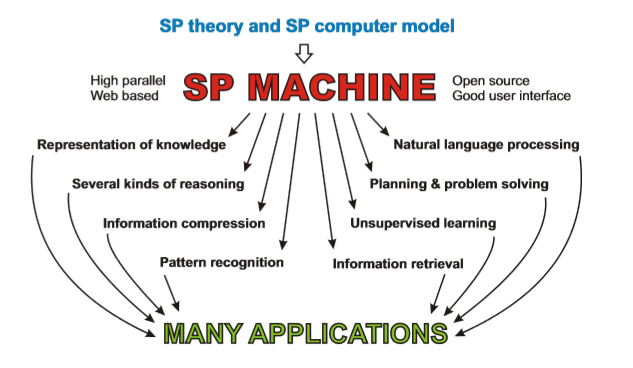
Coding Language : Java (JDK 1.7)

Eclipse tool : Eclipse Indigo

**APPLICATION OF THE PROJECT**

**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.



The SP theory of intelligence combines conceptual simplicity with descriptive and explanatory power in several areas, including concepts of “computing”, the representation of knowledge, natural language processing, pattern recognition, several kinds of reasoning, the storage and retrieval of information, planning and problem solving, unsupervised learning, information compression, neuroscience, and human perception and cognition. In the SP machine there is potential for the simplification of computing systems, including software, with corresponding savings in the time, effort and cost in the development of applications, and other beneﬁts related to information compression.

**EXPECTED OUTCOMES**

* Making big data smaller.
* Analysis of streaming data.
* The unsupervised learning or discovery of ‘natural’ structures in data.

**CONCLUSION**

* The SP system, 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 kind 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. It lends itself to the analysis of streaming data, helping to overcome the problem of velocity in big data.
* the great diversity of formalisms and formats for knowledge, and how they are processed.

**REFERENCES**

J. G. Wolff, ``The SP theory of intelligence: An overview,'' *Information*, vol. 4, no. 3, pp. 283 341, 2013.

J. G. Wolff, National Research Council, *Frontiers in Massive Data Analysis*. Washington, DC, USA: National Academies Press, 2013.

J. G. Wolff, ``The SP theory of intelligence: Benefits and applications,'' *Information*, vol. 5, no. 1, pp. 1 27, 2014

J. G. Wolff, “Application of the SP theory of intelligence to the under-standing of natural vision and the development of computer vision,'' 2013, in preparation.

J. G. Wolff, ``The SP theory of intelligence: An introduction,'' unpublished, Dec. 2013