**AVAILABLE TECHNIQUES AND THEIR LIMITATIONS**

Some of the techniques that are available in the present scenario for converting unstructured data to the structured data are

1. Natural Language Processing
2. Information Retrieval
3. Text Analytics
4. Big data analytics

**1. Natural Language Processing**

Natural language processing is a process of extracting meaning from the text written in the natural language [1].

Some of the natural language processing techniques are [1]

|  |  |  |
| --- | --- | --- |
| **Method** | **Concept** | **Limitations** |
| Stemming | Process of removing the suffixes from the word, this derives a word to their root [1, 4]. | * Over stemming is where the two different words are stemmed to the same root [5]. * Under stemming is where the two different words should be stemmed to the same root but not [5]. |
| Abbreviation Expansion | The abbreviation expressions are expanded to full words [1]. |  |
| Text summarization | Summary in natural language from a larger text [1]. |  |
| Auto completion | Completion of words by predicting the sequence of words [1]. |  |
| Part-of-speech tagging | Tagging the part of speech for each word such as noun, verb, adverb, etc. [1]. | * Removing ambiguities from the context [8]. |
| Sentiment analysis | Predicting the opinion of the words [1]. | * Context – A decision cannot be made based on the words that are used in the context, as there will be two different meaning for the same word [6]. * Regional Variations – Language used in the context is a major limitation as different language words have different meaning [6]. |
| Topic segmentation | Separating the text in to the specific topic [1]. |  |
| Optical character recognition (OCR) | Extracting information from the image [1]. | * Mixture of text and graphics [7] * Use of subscripts and superscripts [7] |

**2. Information Retrieval:**

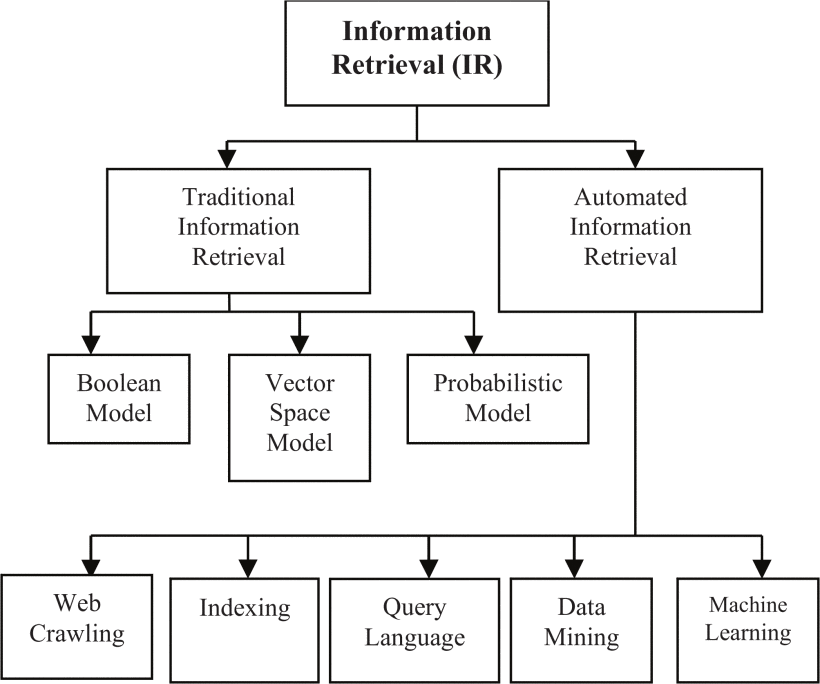
Information retrieval is one of the data mining technique [1] available for converting unstructured data to structured data.

Some of the available IR systems are SMART, Lucene, Lemur, JFeret and Inquire [2]

**Limitations of IR approach**

It doesn’t look for the exact string matching and it is difficult to draw a conclusion from the words used in the data [1].

**Types of Information Retrieval**

****

# Source: [V.N. Gudivada](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.N.%20Gudivada.QT.&newsearch=true), [W.I. Grosky](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.W.I.%20Grosky.QT.&newsearch=true), [V.V. Raghavan](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.V.%20Raghavan.QT.&newsearch=true), Information retrieval on the World Wide Web,  [IEEE Internet Computing](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/RecentIssue.jsp?punumber=4236), 06 August 2002

**Traditional Information Retrieval [2]**

|  |  |  |
| --- | --- | --- |
| **Method** | **Concept** | **Limitations** |
| **Boolean Model** | Boolean model in information retrieval is based on Boolean algebra. It is one of the crucial model of the IR systems [2]. In Boolean model the documents are represented by set of index terms which are viewed as Boolean variable and it is valued as True if the term is present in the document. The queries to retrieve information from the document are expressed in Boolean variables by combining with the logical operators such as AND, OR, NOT [3]. | * Retrieval performance is very poor [3]. * If the user specifies a certain limit of queries by connecting with logical connector AND the results are not retrieved for all the queries [3]. * Retrieval status value is almost same for all the information retrieved [3]. |
| **Probabilistic Model** | The probabilistic model is based upon the probability ranking. The ranking of the document is according to the relevance of query [2]. | * Estimating the term occurrence probabilities in the relevant and irrelevant parts of the document is very difficult [3]. |
| **Vector Space Model** | Vector space model is an algebraic model where each document is represented in vector of terms [2]. | * It is difficult for the vector space model to conclude the words resembling the same concept [1]. |

**Automated Information retrieval [2]**

|  |  |  |
| --- | --- | --- |
| **Method** | **Concept** | **Limitations** |
| Web Crawling | Automatic script that can download the contents from the web pages [2]. | * Content selection is very poor sometimes it bypasses the low-quality content, irrelevant and redundant content [9]. * Some content suppliers inject unwanted content to the site which results in misdirection [9]. |
| Indexing | Process of extracting information rather than analyzing the content [2]. |  |
| Query language | Process of retrieving information from the database by passing the query into them [2]. |  |
| Data mining | Extracting useful information from the huge database process is known as data mining [2]. | * Problems may arise with the missing, corrupted or non-standardized data [10]. |
| Machine learning |  |  |

**3. Text analytics**

Text analytics helps in retrieving the valuable text information from the unstructured and semi-structured data. Text analytics is also referred as text mining [11].

**Steps involved in text analytics:**

1. Retrieving documents for analysis
2. Apply statistical or linguistic or structural technique for further extracting entities, tagging, identifying relationships in the document sets.
3. Apply statistical pattern matching and similarity techniques for classifying document to the specific category.

Text analytics also utilizes techniques like Information Retrieval and Natural language processing for processing and extracting text from the unstructured data sets.

**Limitations of Text analytics**

1. Lot of software programming is needed to extract textual information from different sources.
2. Managing the data unstructured data from various sources is complicated.

**4. Big data analytics**

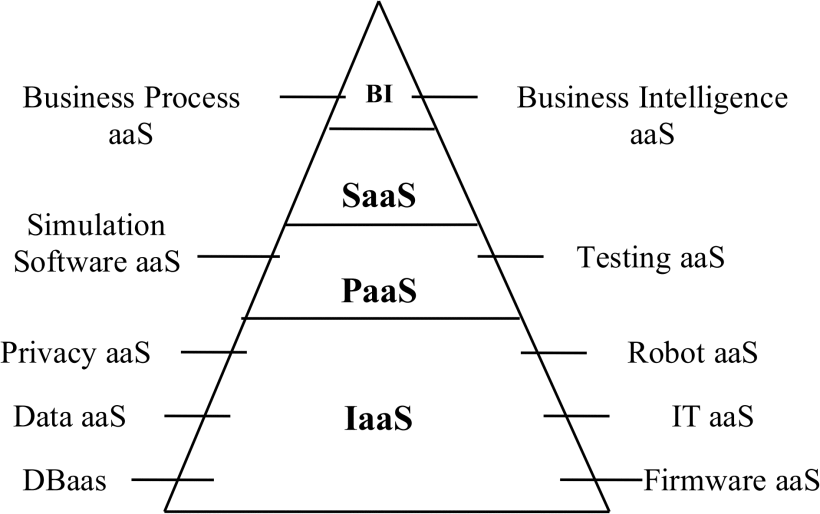
Big data refers to large or complex data that are difficult to manage with the traditional software, hardware and data management tools [12].

Big data is characterized by “3 V’s” [13]

* Velocity – speed of processing the data in a response time
* Volume – refers to the large volume of data
* Variety – refers to the different variety such as structured, unstructured and semi-structured

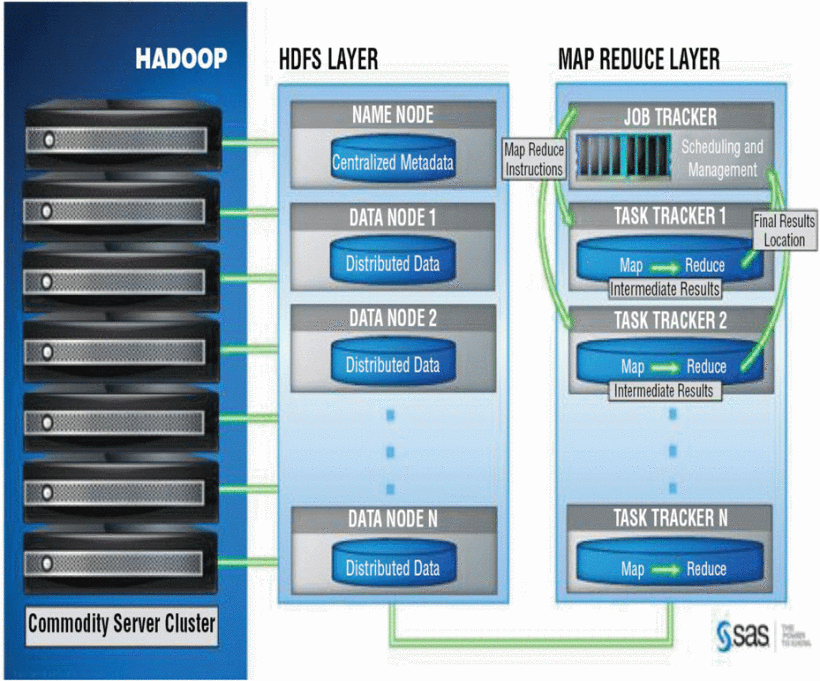
Technologies available in Big data analytics for converting the unstructured data to structured data are [13]

* Cloud computing services helps in supporting and handling big data. It also helps in the major issues like storing and computing the big data. Some resources can be added in cloud computing at any time. The categories in the cloud computing are classified into four and they are Infrastructure as a Service(IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), Business Intelligence (BI).
  + IaaS model helps big data in storing the data from where the data can be retrieved and edited by the customers.
  + The role of PaaS is less than the role of IaaS as there are some limitations for the user as they can work only on the cloud vendor frameworks. From this service layer some of the languages cannot be accessed.



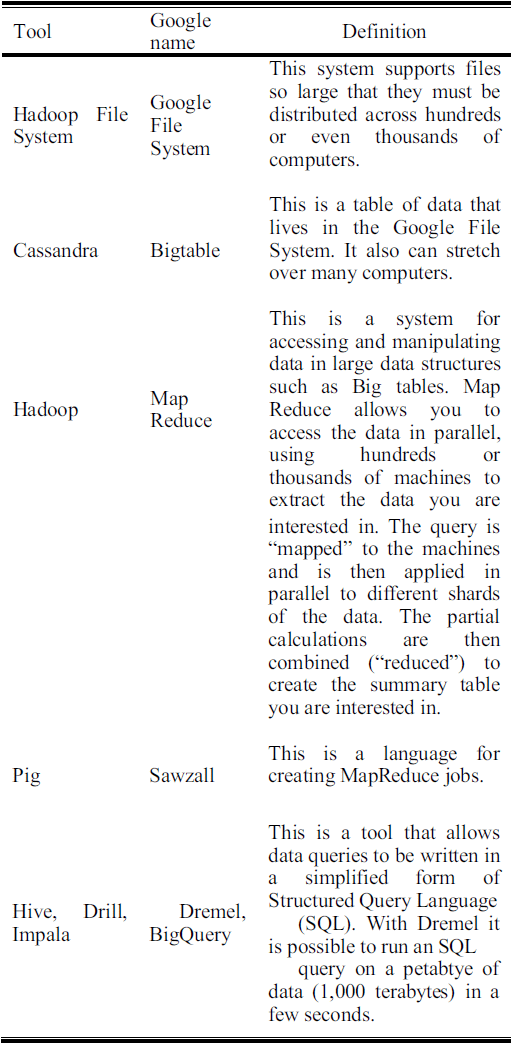
# Source: Hu Shuijing, Big Data Analytics: Key Technologies and Challenges, [Robots & Intelligent System (ICRIS), 2016 International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7754734) 27-28 Aug. 2016

* Hydoop
  + Hydoop is a framework for running application built on large cluster of hardware. The framework in Hadoop provides reliability and data motion. Hadoop provides Map/Reduce where the application can be divided into many small fragments and it can be executed in any cluster. Hadoop also provides distributed file system (HDFS) that stores the data on the cluster nodes with the high bandwidth.
  + Hadoop can be used to store and manage large amount of data and the hardware can be added easily.

****

# Source: Hu Shuijing, Big Data Analytics: Key Technologies and Challenges, [Robots & Intelligent System (ICRIS), 2016 International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7754734) 27-28 Aug. 2016

* **Other tools to process big data**



**Limitations of Big data**

1. Privacy and Security issues
2. Accessibility

**REFERENCES**

1. [Gabriele Bavota](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Gabriele%20Bavota.QT.&newsearch=true), Mining Unstructured Data in Software Repositories: Current and Future Trends, [Software Analysis, Evolution, and Reengineering (SANER), 2016 IEEE 23rd International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7476313) 14-18 March 2016

# [Chandni Saini](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Chandni%20Saini.QT.&newsearch=true), [Vinay Arora](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Vinay%20Arora.QT.&newsearch=true), Information retrieval in web crawling: A survey,  [Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7592392) 21-24 Sept 2016

# [V.N. Gudivada](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.N.%20Gudivada.QT.&newsearch=true), [W.I. Grosky](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.W.I.%20Grosky.QT.&newsearch=true), [V.V. Raghavan](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.V.%20Raghavan.QT.&newsearch=true), Information retrieval on the World Wide Web,  [IEEE Internet Computing](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/RecentIssue.jsp?punumber=4236), 06 August 2002

# [Rahul S. Dudhabaware](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Rahul%20S.%20Dudhabaware.QT.&newsearch=true), [Mangala S. Madankar](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Mangala%20S.%20Madankar.QT.&newsearch=true), Review on Natural Language Processing Tasks for Text Documents,  [Computational Intelligence and Computing Research (ICCIC), 2014 IEEE International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7203316), 18-20 Dec. 2014

# [Cristian Moral](http://www.informationr.net/ir/19-1/paper605.html#author), [Angélica de Antonio](http://www.informationr.net/ir/19-1/paper605.html#author), [Ricardo Imbert](http://www.informationr.net/ir/19-1/paper605.html#author) and [Jaime Ramírez](http://www.informationr.net/ir/19-1/paper605.html#author), A survey of stemming algorithms in information retrieval, March, 2014

# [V Nagarjana Devi Duvvuri](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V%20Nagarjana%20Devi%20Duvvuri.QT.&newsearch=true), [V.Rajanikanth Thatiparthi](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.V.Rajanikanth%20Thatiparthi.QT.&newsearch=true), [Rajashekar Pantangi](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Rajashekar%20Pantangi.QT.&newsearch=true), [Akhil Gangavarapu](http://ieeexplore.ieee.org.ezproxy.gannon.edu/search/searchresult.jsp?searchWithin=%22Authors%22:.QT.Akhil%20Gangavarapu.QT.&newsearch=true), Sentiment Analysis Using Harn Algorithm, [IT Convergence and Security (ICITCS), 2016 6th International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7740253) 26-26 Sept. 2016

# Line Eikvil, Optical Character recognition, December 1993

# William M. Darling, Michael J. Paul, Fei Song, Unsupervised Part-of-Speech Tagging in Noisy and Esoteric Domains with a Syntactic-Semantic Bayesian HMM

# Namrata H.S, B.S. Satpute, Pramod Patil, Web Forum Crawling Techniques, International Journal of Computer Applications (0975 – 8887) Volume 85 – No 17, January 2014

# Hian Chye Koh, Gerald Tan, Data Mining Applications in Healthcare

# Kalli Srinivasa Nageswara Prasad, Prof S. Ramakrishna, Text Analytics to Data Warehousing.

# Wullianallur Raghupathi and Viju Raghupathi, Big data analytics in healthcare: promise and potential, 07 Feb 2014

# Hu Shuijing, Big Data Analytics: Key Technologies and Challenges, [Robots & Intelligent System (ICRIS), 2016 International Conference on](http://ieeexplore.ieee.org.ezproxy.gannon.edu/xpl/mostRecentIssue.jsp?punumber=7754734) 27-28 Aug. 2016