2. LITERATURE SURVEY

2.1 INTRODUCTION

Semantic Content Management System in cloud recently captured general consideration from researchers across many domain-specific applications. This chapter delivers a detailed discussion of literature related to various technical challenges in Cloud, CMS and Semantic Content Management Systems. It also deals with the functionalities such as Information Retrieval in current web search engines and the issues in the existing semantic similarity measures, especially ontology-based measures. This chapter analyses and presents a detailed insight into each of the existing methods and discusses the unique aspects of the proposed DSCMS model from the existing ones.

This thesis aims at developing a DSCMS which is deployed in the cloud. This model is developed with various information retrieval techniques implemented in the distributed framework. The benefit of this DSCMS model is threefold. Firstly, it provides the cloud users with data extraction, searching and retrieval of semantic content which is purely concept-based retrieval and not keyword-based retrieval. Secondly, the analysis engine which comprises of rendering engine with the machine learning algorithms are developed and implemented and finally, the inbuilt tracking system is incorporated along with a recommendation system for detecting the changes dynamically in the weblogs.

2.2 CLOUD COMPUTING AND SaaS – AN OVERVIEW

Cloud is a computing technology that involves a pool of computing resources bound together with a more robust internetworking communication. A thin client is more than enough to handle the remote virtual hardware and software resources. Software service delivery model is hosted centrally and it is “Pay per use software”. SaaS has become a regular delivery model online which works via thin clients. Content Management System is modeled to be SaaS. The Proposed DSCMS is also deployed in the cloud as a service for
the cloud users. In [26] sharing media between the home networks is described and implemented. The content has been managed in all forms like images, videos and text. The accessing resources across the internet using the model are cost-effective compared to the existing ones which are very advantageous. The process implemented in this model does not involve semantic functionality. Even then the model is inspiring for deploying the DSCMS in the cloud which can be cost-effective and efficient. In WFCMS, web based content management system introduces systematic work principle, property and application. The MVC architecture and implementation were described with performance evaluation. Enterprise CMS is similar kind of web CMS [27], it provides functionality to control and analyze information. Key challenges in ECMS are identified and a road map for execution was well defined. The paper also insists that the inclusion of new functionalities is also still open for research.

There has been lots of research in cloud-based CMS and semantic Information Retrieval in documents. Some of the works have been gathered to create a model which helps the users to manage and rank the sources for their future usage. The research study [28] presented a new user interface for semantic authoring. Semantic Content Authoring has been proposed. Four types of SCA have been explored with the elaborate explanation for instuitive ideas for development of novel semantic authoring interface. Quality attributes have been analyzed for the SCA, a semantic rich content has been created semi-automatically by means of this method. Storage and further usage of SCA is not implemented with the aspects of collaboration, user-friendliness and automation.

A semantic web search is presented in [29] where the queries are processed. It is based on rules of reasoning. A standalone model with the desktop search with page ranking is generalized. The search has been enabled with movie data store.

The web search and retrieval for a semantic user interface is well defined in [30] which proposed a semantic web search using Natural Language Processing (NLP). A complete end to end process of
pre-processing, annotation, semantic analysis and executing a SPARQL query is implemented in the paper. The inconsistencies of web data and usage of semantic web framework is justified by the author. The limitations include analysis of graph set and ontology based analysis. Ontology plays a major in semantic web, hence in addition to these models there is a need for literature survey in ontology and its features in the semantic web.

In [31] the authors presented a web OWL, which is used to construct a search engine. The system is implemented using Jena API and has captured and saved more of class and properties. Similarly, in [32] S-trans method is proposed to conversion of XML schema to OWL. The schema was applied to a health care data integrated into a different domain. This paper gave an idea to convert the relationship and rules to RDF or OWL ontology.

Another ontology-based application [33] has thrown light on online intelligent applications and digital libraries. The problem of ontology has been addressed and CORESE was proposed. The authors evaluated the proposed systems from the end-user view point and applications perspective. Model has been evaluated on a schema based on advanced XML based RDF.

Authoring and the user interface design are necessary for any web search engines and also CMS, to throw light on those areas [34] of semantic authoring. An intelligent user interface namely CHOCOLATO has been developed with the rich semantic interface. It guides the students and the teachers to use the system effectively. Accordingly [35] presented a Thesus model. It is a language based system which uses integration of linked data. The linked data in semantical analysis is concentrated and demonstrated with the evaluations.

2.3 INFORMATION RETRIEVAL BASED ON SEMANTIC ANALYSIS

Above discussions in research studies show that there are various semantic techniques and models developed and implemented in Information retrieval. But there are few challenges inferred from the literature based on
semantic web search and information retrieval from CMS deployed in the cloud. The problems are listed below:

- Lack of automatic semantic techniques for extending the work to manage complex queries in the cloud.
- Though many researchers had defined the semantic ontology for query retrieval semantic similarity score based on ontology is limited and not very much accurate.
- Information retrieval techniques are very much older and conventional when performed with keyword matching. There is a lack of defined content related matching of documents.
- Automated locating and tracking mechanisms are not identified and implemented for the modern collaborative and semantic web. The changes in the sources are not detected and also the changes are not intimated to the users. The functionalities that are incorporated in the SCMS are similar to semantic web search engines. It includes processes such as:
  - Data extraction and semantic crawler
  - Storage in repository
  - Semantic Clustering using Map-Reduce paradigm
  - Semantic Classification and ranking
  - Recommendation using semantic similarity measure
  - Searching and querying engine for change detection.

The survey about the above functionalities is described in the following sections.

### 2.3.1 Data Extraction using Semantic Crawler and Storage

In this section, literature for extraction of the data from different search engines and semantic relatedness is explored. The usage of semantic [36] crawler and its storage in the repository is discussed. Data extraction is a tedious task for unstructured data. The cloud model is accessed as pay per use model. The relational databases are used to save schema based data.
But the requirement of unstructured data is entirely different, so there is a need for an unstructured database system which is identified as document-oriented storage in the proposed research.

Document Oriented Database, is a scalable, schema-less system, high-performance, these databases are suitable for services on cloud. Both the technologies are popular now which overcomes the drawbacks of conventional systems. The disadvantages of distributed computing and infrastructure are avoided by on demand based applications. Issues in metadata extraction, semantic relatedness, semantic crawler and storage are discussed in the following literature.

The research study in [37] proposed a semantic system which works on world news domain. The system relies on the meta-data based on the script file. When the user triggers a query, the metadata files are used for searching rather search from the whole document. The above method is achieved by World News Ontology and heuristic methods confined to specific domain are developed.

The research proposal in [38] developed resources in the semantic network are defined and the relation between the resources are taken in to account. It also identifies the relevant information based on ontology. The scores based on weight are assigned to the web pages. Based on the weights assigned, ranking is performed and also uniqueness of resources is maintained. The work in [39] proposed a system which matches the words based on semantic similarity. Matching of the semantically related words and documents are expressed in this paper. The evaluation is conducted by the manual identification for datasets which gives solution for choice problem. The retrieval can be very well accomplished by reasoning with ontologies which is incorporated in [40] which proposed a novel approach to find similar entities with linkage of data. The proposal combines multiple lexical matching strategies by applying a summarization technique; then it employs anatomical data and similarities to find latest data.
The literature for retrieval and matching of the content with the existing ones are discussed till now, but for possible pairing, there is a need for efficient storage. Storage can be accomplished by the content stored by crawlers. Crawler harvests the URL and stores the content in the repositories. Two different semantic crawlers are explored. In [41], priority based web crawler algorithm has been proposed and ontology in semantic web is used to crawl the content. A priority queue is used for implementing and the performance is evaluated with the existing simple crawler. In [42], semantic metadata is used for matching the content with the crawled web site. This model promoted proper information storage with the help of ontology and illustrated the benefit of the semantic crawler for the updated scope of the search engine.

In addition to this in [43], the storage in Non-Relational databases is discussed with examples. The streaming and growing content cannot be stored in the traditional databases. The process of online analytical processing and need for different types of NoSQL are dealt in this paper with features and working. This paper assisted in exploring an improved storage mechanism.

As a result of reviews from above all papers it is concluded that the semantic analysis of documents in the web yield more accurate retrieval results. There were also discussions about storage of the web documents in a suitable repository. The meaningful storage is provided by the semantic crawlers. Hence there is a discussion about semantic crawlers. Meta content extraction and storage plays a significant role in the content management systems; the content reviewed here shows that it is essential to incorporate semantic crawler for meaningful storage. Efficient storage can be established by a grouping of documents by clustering and classification method which is dealt in next section.

2.3.2 Semantic Clustering of Documents

The clustering is one of the main branches of data mining especially text mining, the grouping of data or objects is based on the similarity between
the data. The clustering is performed with the automatic machine learning algorithm dynamically. The signature of the data can be obtained from the clustered group. The pros and cons in clustering using map-reduce paradigm are discussed.

In the literature [44], the clustering of XML documents with the XML query and there are two important methods in the paper, in which traditional method is used for search and retrieval, whereas the other method performs the processing dynamically. The interests of users are recorded and it is segregates with different granularities to locate the effects quickly. The authors have also demonstrated the results with justification.

The scalable content can be handled with the Map-Reduce Paradigm. Clustering using map-reduce is the tedious task which is accomplished in [45], this paper compares K-means and fuzzy C-means which is implemented on noisy and genuine big dataset. In huge dataset, there is an obvious difference between the execution time of both the algorithms. The mahout based algorithm on the wiki data set. From these papers, it is presumed that algorithm is not faster. In [46], topic modeling based on latent semantics is discussed. This method also assists document clustering.

The discussions and implementations in [47] explain the types of semantic document clustering. The drawbacks in the conventional vector space models are discussed in detail. A semantic approach with calculated similarity score is utilized for matching the text documents with the existing topics. But these types of clustering are not performed in Map-Reduce paradigm.

The paper [48] [49] focuses on combination of unsupervised and supervised algorithms to perform document clustering. A fuzzy K-means algorithm is introduced to compute distance between the data. The author conducts the investigations on three different datasets. The experimental results indicate that fuzzy semi-K-means can outperform the other methods.
In paper [50], the document clustering is performed based on semantic understanding. The similarity between two web documents is calculated and measures are identified from the semantic marks form the file. The semantic methods have enabled more efficient document clustering technique than the traditional methods.

The above literature illustrates clustering without the map-reduce paradigm. There is a need to review the research papers based on map-reduce paradigm using ontology and RDF. Since the proposed system is implemented based on a distributed framework with the map-reduce paradigm.

In the paper [51], the design of scientific workflows is processed through XML pipelines are transformed to map-reduce network flows. The work demonstrates a considerable reduction in the execution time. Specific compilation strategies are designed and executed in the system. This paper contributes the idea of reducing the uploading, searching, retrieval time of the proposed thesis. The work [52] illustrates the working of a map-reduce model in large-scale handling ontologies. The growing ontologies and triples from the RDF are handled by map reduce and results in less execution time. This paper supplements the above idea of the proposed thesis.

The literature of the above works, it is obvious that semantic analysis fetch accurate results from the retrieval. In the proposed work clustering is performed semantically in distributed framework using machine learning algorithms. It provides precise retrieval outcomes compared to conventional methods. When clustering is implemented with the map-reduce paradigm, it reduces the execution time to a greater extent. Clustering is an unsupervised technique and the labels of the data sets are given as inputs to the classifier labels. Classification is a kind of supervised learning which earns more accurate results when implemented with a combination of unsupervised learning.
2.3.3 Semantic Classification of Documents using Ranking Technique

Classification is one of the types of supervised learning; the labels of data considered for classification are well defined. The threshold values for categorizing the documents are analyzed in the clustering technique using semantic tool. Moreover, a massive amount of documents in the search engine are ranked using page ranking technique based on classifiers with a threshold value. The issues in classification techniques and ranking are discussed in the following sections. There are two types of classifiers considered in the thesis, like content-based classification and score based classification hence the literature survey is as follows.

In the paper [53], automatic keyword extraction, automatic summarization is the critical focus. The keyword extraction is enhanced with collaborative bagging methodology with random forest algorithm. The f-measure of this method proved to be better than other conventional algorithms. The Naïve Bayesian algorithm is also used along with support vector machine, logistic regression and random forest with the widely used ensemble methods. This paper inspired in determining four-fold similarity in the proposed research.

The paper [54] describes the cross-language WikiBoc methodology used for classifying the content based on the concept. Concept matching is a significant focus in this method. Several experiments are conducted with three lingual corpora. Algorithms have been written for extracting the concepts from the content. These concepts are given as input to the classifiers. This kind of classifier outperforms the existing classifiers. Content level classification is not enough for ranking the documents; the gain has to be calculated based on scoring strategies. Literature-based on scoring and classification meant for scoring methodologies should be necessarily reviewed.

Ranking of documents in the web is performed with the help of page ranking algorithms and hub score methodologies. In [55], hub score algorithm is used. Inter-ontology and intra-ontology are compared for identifying the
authoritative and centrality scores. Mathematical formulation shows the probability of detecting the best scores for a particular document with the help of relevancy. The page ranking algorithm is implemented to rank the web pages and effectiveness of algorithm is enriched with dual walk rank algorithms. Extraction of the implicit ontology links is implemented with an explicit find relevancy algorithm. The limitations like consideration of concept matching in the creation of ontology are succeeded in the proposed work.

Score level fusion approach based on the classification is demonstrated in work [56], two different algorithms such as sentence based vector space model and neural network based representation of documents. The score calculated from both the algorithms are improvised using fusion score approach. In the paper [57], Hadoop updated (HadUp) and tailored Hadoop architecture is demonstrated. The paper uses differential algorithm which matches the old and new datasets and eliminates the unwanted. Less computation power is enough to process the data sets. This method indicates that classification can very well be implemented in Hadoop.

The paper [58] insists the use of semantic ontology in the news domain. The knowledge management in news domain is a big challenge. The production and distribution of news can be enhanced based on the news ontology. A protégé extension was developed to automate the process partially. A distributed network is used to handle the data in this paper. The distributed network is refined and map-reduce paradigm models are introduced in the proposed system to overcome the limitations.

From the above conclusions of existing work, it is clear that semantic analysis applied to any document or system yields valid results. There are only limited semantic oriented machine learning algorithms in distributed system. In the current work, hub score based ranking and recommendation are implemented with filtering algorithms.
2.3.4 Recommendation based on Semantic Similarity

Recommendation system comprises of features, content that defines a document or a user. It also involves similar users or overlapping users. This system can be enhanced by the semantic similarity approach among the features. In case of documents, the synonyms of the selected text are compared to the benchmark test. The meaning of the benchmarked text is already known. The dialogue-based intelligent tutoring systems are demonstrated. Recommendation of best answers based on word sense similarity is implicated based on the similarity approach. Searching based literature and recommendation methods are explored in this section.

Searching and retrieval of the desired content involve recommendation and specific querying mechanisms. The Linked Open Data (LOD) is enhanced by the semantic content. LOD is also used for advanced searching and querying mechanisms provided by SPARQL. The semantic analysis in LOD has paved the way for innovative research not only for improving existing applications but also for realizing innovative semantic applications [59]. The author has demonstrated various functionalities in querying, but it is inadequate to deal with features such as comparing, ordering and ranking search results. These processes are fundamental to applications such as recommendation provision and matchmaking. The work in [60] explored the features of the medical systems that need resources to handle the data. However, it leads to limitations of Health Sciences. Adding experts dynamically in database needs a scalable storage. This kind of growing database needs a distributed framework.

In the paper [61], the author described Berlin SPARQL Benchmark (BSBM) and a comparative analysis is accomplished with native RDF stores and SPARQL-to-SQL rewriters across architectures. The limitations faced in this paper were SPARQL is still a very new query language it is likely that the RDF stores have not yet implemented in the distributed framework. There are no models based on RDF stores in DFS. Personalization is performed as one of the significant feature in recommendation system. The user centric algorithm for the filtering the results are proposed in [62]. In the user centric
filtering algorithm recommendation is introduced based in input as well as user preferences collaboratively.

Accompanying the above work Keyword Aware Service Recommendation system (KASR) is proposed and demonstrated in [63], in this system keyword-based querying and analysis is executed over map-reduce paradigm. The retrieval is remarkably faster. The precision is considerable when evaluated with the metrics. The performance is improved by map-reduce techniques. The limitations such as semantic analysis in KASR system is overcome in the proposed system.

The proposed HBSRS overcomes the above limitations in classification techniques and querying methodologies in this thesis. The inbuilt tracking system is necessary for the recommended URLs.

2.3.5 Searching and Query Engine for Change Detection

Managing content over a distributed database system reduces the processing time and comes with better performance and accuracy. But the information retrieved sometimes contains irrelevant files which are not based requirements of the end user. Querying and searching for the required content in such data store is a tedious task. Semantic searching and retrieval based on RDF triples are accomplished by processing it. The output of the searched web documents is recommended. The recommended documents are monitored from time in the cloud using feeds based on Really Simple Syndication (RSS) and atom in the wiki. The papers contributing to RDF and change detection are reviewed in the following sections.

The author of the paper [64] introduces the triples based search, in which the query sent by the user is processed based on the depth decision in the triples hierarchical tree. The relationship between the subject, predicate and objects is depicted via the triples based trees, which updates the objects easily. The performance is improved by this approach. The work is limited in terms of dynamic updating of triples in the tree.
In [65], wangchao et al., the summarization of RDF data makes the search efficiently. The billions and trillions of RDF data are handled by the system proposed. The time taken by this system is more when compared to distributed system.

The keyword search is very efficient for the schema based data in [66], the retrieval is based on the concept of sub-graphs. The algorithm based on the retrieval uses sub-graphs and also the ranking techniques. The author in [67] presented a novel platform called blog connect. The blog combines the tracking code integrated into newly developed platform called Blog Connect. This blog connect system query’s the system and the results are retrieved based on the co-keyword of the keyword. The complexity of the system is reduced by removing redundant documents based on the secondary keyword.

The tracking of existing web pages and identifying the changes were proposed by [68], the frequent changes are easily detected for specific parts in the document. The change detection for recommended web URLs is accomplished in the proposed system. It is performed for monitoring the structural and content changes within the particular web source document.

In [69], the change detection technique in a focused web crawler is dealt in this paper. The collections of downloaded web pages are maintained and if the page has to be tracked, it is revisited in the regular intervals. Architecture has been proposed to compare the old URL and new URL. The authors have tested one hundred pages and there were only changes in 12 pages out of the test bed considered.

In [70], the authorization framework for the collaborative cloud is implemented. The access control policies for the different users in the system are well defined. Based on the above work the access control policies are established for the users in the change tracker in the proposed DSCMS. The above literature concludes that the semantic techniques have been proposed for improvising web search engines and evaluating complex queries. Also it
is used to identify the changes in the source document after it has been referred and stored.

In nutshell, the white space that has to be addressed in each of the components in the proposed system with relevant parameters is given below.

The fourfold semantic similarity in DSCMS is combined from all three phases like Data extraction & Semantic crawling, Semantic Rendering Engine which comprises of Semantic Clustering, Semantic Classification and Ranking, Semantic Recommendation & Inbuilt change tracker implemented in Hadoop Distributed File System (HDFS). The gap in the paper [71] is addressed by improvising the design of semantic CMS with all the above mentioned components.

The three fold semantic similarity measure which is used as semantic signature is addressed in [118] has been improvised in the proposed research.

In the existing semantic crawlers only Meta-keyword based seed URL harvesting is proposed and there is a need for structure and domain based Semantic Structure and Meta-keyword Crawler which can improve the Harvest Rate [89] & [90].

In existing literature K-Means clustering and Hierarchical Agglomerative Clustering are implemented in Hadoop distributed file system [108], [109]. It is required to improvise the clusters in map reduce semantically so that there will be increase in accuracy of cluster formation and decrease in time.

Semantic classification is implemented with known labels in mapper and reducer based on distances calculated from ontological triples inside Hadoop [55], [113] and ranking techniques. There is a gap to increase Recall, Precision and F-measure by using the score based or sequence words based classifiers. Key word based recommendation system using collaborative filtering is implemented [63]. It is still necessary to design Content based semantic recommendation system for accurate retrieval.
Change detection is available based on change logs, URL keywords and timestamp but the content of the URL based on the structure of web documents and updating of triples is still a challenge [65], [68]. And it is also necessary to improve the percentage of change detection based on the content which will benefit the e-learners.
2.4 SUMMARY

- In this chapter, need for semantic retrieval system and its methodologies, in distributed storage framework for content management are discussed.

- Various semantic models are explored, semantic techniques and algorithms incorporated in the models are studied. The involvement of many research groups for semantic content retrieval in a cloud environment is identified and the limitations of those studies are reviewed.

- The techniques incorporated in data extraction, semantic crawling and storage of data are presented. The weaknesses in the existing crawlers and storage techniques are discussed. There is a lack of user focused semantic crawler with distributed databases.

- The review also presented through survey of clustering, classification and recommendation systems. It insists the need for a combination of unsupervised and supervised learning methods. The review of ranking and recommendation system shows that it enriches the accuracy to a greater extent.

- The need for the inbuilt tracking system and semantic analysis is discussed with the conventional change detector algorithms.

- The comprehension of white space that has to be addressed with the relevant parameters and the challenges are also mentioned.

- In this work, the shortfalls and limitations of the SCMS and various methods used in it are reviewed meticulously and a DSCMS model is proposed with the problem definition.

The next chapter illustrates the architecture of the proposed model with the detailed explanation about the components.