ABSTRACT

In today’s digital era, cloud computing gained popularity in providing demand-based services through various service models. Software as a Service (SaaS) is one of the prominent service model in cloud. With the advent of web collaborations and social media, massive number of documents are created, managed and published. Thousands of users are collaborating and contributing to the Content Management Systems (CMS) like Wikipedia and weblogs. CMS is identified as one of the application in SaaS. Existing CMS make a template of the content, perform authoring and then publish the content without any semantic analysis. Linked data and sources for articles are annotated with a massive number of references. But the relevant, updated and exact sources are not annotated semantically in existing CMS which leads to conflict in contents. Moreover, there is no inbuilt change tracking system along with the CMS to update and track the changes of the refereed sources. Addressing the above issues, the research aims to develop a Document based Semantic Content Management System (DSCMS) model in the cloud which efficiently crawls and manages the stored documents, retrieve the relevant recommended documents and track recommended source web pages. The DSCMS model is a three-tier architecture model which consists of a Semantic Content Marker (SCM) with Semantic Crawler, Rendering Engine for Semantic Clusters and Classification (RESCC) and Hadoop Based Semantic Recommendation System (HBSRS) with Inbuilt Change tracker System (ICTS).

The end user retrieves the recommended and relevant URLs. Here, semantic relevancy is achieved by introducing fourfold semantic similarity measure such as word sense relativity, domain-based relevancy, sequence-based word sensitivity and content-based relationship in all the components of the proposed model. When the end user searches for a new topic, the semantic crawler harvests the relevant source web documents. Extraction, transformation and loading are facilitated in this research by SCM which uses concept matching and semantic annotation techniques. When the end user searches already existing topic, the semantic retrieval is accomplished. In
RESCC, there are considerable amount of linked, annotated documents that are to be stored and rendered. This work is realized by domain expert users in which semantic clusters are generated using preliminary semantic K-means algorithm. The accuracy of the cluster is improvised by proposing a mapper and reducer for Semantic Hierarchical Agglomerative Clustering (SHAC). SHAC proved to be the best solution for clustering the documents which enable faster searching and retrieval. Ranking of most relevant documents is realized by Semantic Dual Walk ranking algorithm (SDWR) which retrieves the appropriate documents with greater precision. The results from the SHAC are applied as labels to classification techniques. Sequence-based Naïve Bayesian (SNB) classifier is proposed which resolves the issue of polysemy (same word interpreted in different meaning in other field or domain) occurrences of normal words and disambiguate words. The classified web documents are recommended in the next phase.

HBSRS implements User Focused Blog Crawling (UFBC) algorithm and Collaborative Filtering (CF) techniques to personalize the recommended results. Automated Breath First Search (ABFS) is proposed in ICTS to compare and track the structural and content changes in the source web pages. It periodically alerts the user by implementing Dynamic Streaming Based Change detection (DSBC) scripts. DSCMS is implemented by Semantic Retrieval of Source Content (SRSC) model. Quantitative comparison proves that the unrelated links are pruned and only relevant source web documents are retrieved. DSCMS shows considerable increase in accuracy rate and the relevant retrieval rate is approximately increased up to 40%. The system also reduces $1/15^{th}$ of the execution time when compared to existing keyword based content retrieval systems. Qualitative comparison with the current CMS proved that the DSCMS is more proficient regarding functionalities, and it is faster and accurate in the retrieval of relevant documents compared to the conventional CMS. In addition to this, the inbuilt tracker detects the dynamic changes in the collaborative web documents and alerts the user. The fourfold semantic solution provided in the distributed environment make the system more efficient than conventional methods like keyword, semantic and lexical based retrieval.