Abstract—Web Crawlers are one of the most critical components used by the Search Engines to collect pages from the Web. It is an intelligent technique of browsing used by the Search Engine. The requirement of a web crawler that downloads most relevant web pages from large web is still a major challenge in the field of Information Retrieval Systems. Most Web Crawlers use Keywords base approach for retrieving the information from Web. But they retrieve many irrelevant pages as well. Focused crawler is used to collect those web pages that are relevant to a particular topic while filtering out the irrelevant pages.

The Ontology based web crawler algorithm used to mining user queries by using separate service. The use of algorithm like page rank and other importance metrics has scheduled a new approach in prioritizing the URL queue for downloading higher relevant pages. The idea of this paper is based on the design of an unsupervised framework for vocabulary-based ontology learning, and also a hybrid algorithm is used for matching semantically relevant concepts and metadata.

Keywords—Information discovery, Mining service, Ontology learning, SASF crawler.

INTRODUCTION

Crawlers (also known as Robots or Spiders) are tools for assembling Web content locally. Focused crawlers in particular, have been introduced for satisfying the need of individuals (e.g. domain experts) or organizations to create and maintain subject-specific web portals or web document collections locally or for addressing complex information needs (for which a web search would yield no satisfactory results). Applications of focused crawlers also include guiding intelligent agents on the Web for locating specialized information. Typical requirements of such application users are the need for high quality and up-to-date results, while minimizing the amount of resources (time, space and network bandwidth) to carry-out the search task. Focused crawlers try to download as many pages relevant to the subject as they can, while keeping the amount of not relevant pages downloaded to a minimum number.

Crawlers are given a starting set of web pages (seed pages) as their input, extract outgoing links appearing in the seed pages and determine what links to visit next based on certain criteria. Web pages pointed to by these links are downloaded, and those satisfying certain relevance criteria are stored in a local repository. Crawlers continue visiting Web pages until a desired number of pages have been downloaded or until local resources (such as storage) are exhausted.
II. PERSPECTIVE SOLUTION
Hidden Markov Model for fetching relevance and non-relevance web pages is used. And priority crawling scores to Rank and order the relevant URLs for getting the highest priority and lowest priority is used. And finding the Similarity value based on priority values, and getting the optimal threshold value and ontology data.

III. RELATED WORK AND LITERATURE SURVEY
A semantic focused crawler is a software agent that is able to traverse the Web, and retrieve as well as download related Web information on specific topics by means of semantic technologies [1], [2]. Since semantic technologies provide shared knowledge for enhancing the interoperability between heterogeneous components, semantic technologies have been broadly applied in the field of industrial automation [3]–[4]. The goal of semantic focused crawlers is to precisely and efficiently retrieve and download relevant Web information by automatically understanding the semantics underlying the Web information and the semantics underlying the predefined topics. A survey conducted by Dong et al. [5] found that most of the crawlers in this domain make use of ontologies to represent the knowledge underlying topics and Web documents. However, the limitation of the ontology-based semantic focused crawlers is that the crawling performance crucially depends on the quality of ontologies. Furthermore, the quality of ontologies may be affected by two issues. The first issue is that, as it is well known that ontology is the formal representation of specific domain knowledge [6] and ontologies are designed by domain experts, a discrepancy may exist between the domain experts’ understanding of the domain knowledge and the domain knowledge that exists in the real world. The second issue is that knowledge is dynamic and is constantly evolving, compared with relatively static ontologies. These two contradictory situations could lead to the problem that ontologies sometimes cannot represent real-world knowledge, considering the issues of differentiation and dynamism. The reflection of this problem in the field of semantic focused crawling is that the ontologies used by semantic focused crawlers cannot precisely represent the knowledge revealed in Web information, since Web information is mostly created or updated by human users with different knowledge understandings, and human users are efficient learners of new knowledge. The eventual consequence of this problem is reflected in the gradually descending curves in the performance of semantic focused crawlers.

In order to solve the defects in ontologies and maintain or enhance the performance of semantic-focused crawlers, researchers have begun to pay attention to enhancing semantic-focused crawling technologies by integrating them with ontology learning technologies. The goal of ontology learning is to semi-automatically extract facts or patterns from a corpus of data and turn these into machine-readable ontologies [7]. Various techniques have been designed for ontology learning, such as statistics-based techniques, logic-based techniques, etc. These techniques can also be classified into supervised techniques, semi-supervised techniques, and unsupervised techniques from the perspective of learning control. Obviously, ontology-learning-based techniques can be used to solve the issue of semantic-focused crawling, by learning new knowledge from crawled documents and integrating the new knowledge with ontologies in order to constantly refine the ontologies.

IV. PROPOSED WORK
To design a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set. Hidden Markov Model is proposed to determine the relevance and non-relevance of a retrieved web page.

And priority crawling score is used to Rank and order the relevant URLs for getting the highest priority and lowest priority and finding the Similarity value based on priority values. Finally getting the optimal threshold value and ontology data. In order to evaluate the retrieval effectiveness, precision Pr corresponds to the fraction of top r ranked web pages that are relevant to the query over the total number of retrieved web pages. The time spent for crawling and analyzing the semantic web pages is another important element to measure.

V. PROPOSED ARCHITECTURE/PROTOTYPE

VI. SCOPE OF WORK
Need to determine the relevance and no relevance of a retrieved web page. For this purpose proposed the Hidden Markov Model for fetching relevance and non-relevance web pages. And used priority crawling score to Rank and order the relevant URLs for getting the highest priority and lowest priority. And finding the Similarity value based on priority values. Finally getting the optimal threshold value and ontology data.
VII. DISCUSSION

Focused crawler is used to collect those web pages that are relevant to a particular topic while filtering out the irrelevant. Proposed system is intend to design a semi-supervised approach by aggregating the unsupervised approach and the supervised ontology learning-based approach, with the purpose of automatically choosing the optimal threshold values for each concept, while keeping the optimal performance without considering the limitation of the training data set. In the proposed system, we need to determine the relevance and non-relevance of a retrieved web page. For this purpose we propose the Hidden Markov Model for fetching relevance and non-relevance we pages by implementing. And we use priority crawling score to Rank and order the relevant URLs for getting the highest priority and lowest priority to get the optimal threshold value.

VIII. REFERENCES