A Comparative Study on Self Adaptive Semantic Focused Crawler and Novel Focused Cell like Membrane Computing Crawler

S. Gunasekaran, Ph.D.
Head of Computer Science and Engineering Department
Coimbatore Institute of Engineering and Technology
India

M. Anisha
PG Scholar,CSE
Coimbatore Institute of Engineering and Technology
India

P.R. Joe Dhanith
Assistant Professor,CSE
Coimbatore Institute of Engineering and Technology
India.

ABSTRACT
World Wide Web is a repository containing an enormous amount of documents and hyperlinked documents. The information on the web is shared based on the user interest. The search through these documents considers the user query and retrieves documents that are related. Traditionally the search engine includes three operations such as searching, indexing and downloading. Among the three operations downloading is the most important one, where few thousands of web pages are downloading with the respect to the user query. This leads to a known problem called information kill where the user looks up only few results and turns away from others. Thus to avoid problem of downloading lot of web pages several crawlers are designed that improves the efficiency of crawling specific documents. Two crawlers namely Self adaptive Semantic Focussed and Cell Membrane Computing Focussed Crawler has been studied with comparison.

Keywords
Crawlers; Semantic Similarity; Topical Seed; Generic Seed.

1. INTRODUCTION
A crawler is a software program used to create search engine index entries by visiting Web sites. It systematically reads the web pages and retrieves information from those pages for web indexing. Web crawling software is used in web search engines to update the web content of the web sites. Web search engine indexes the downloaded pages by crawler for later usage by the user, making the search quicker. Web Crawler is incessant running programs which download pages at regular intervals from internet (Jaytrilok Choudhary, Devshri Roy, 2013) [4]. For assembling Web content locally, crawlers are used as tools. Web crawlers are used in many applications where large number of pages is quickly fetched into a local repository and is indexed based on keywords. Since crawlers extract information from web sites, they are used in Web Scrapping.

A web crawler collecting web pages that are satisfying the particular property is called Focussed Crawlers. These properties are specified by prioritizing the visited pages and outstanding requests that are stored as crawl frontiers. Crawl frontiers are prioritized by classifiers. Focussed crawlers can also manage the hyperlink exploration process. Before downloading any page a focussed crawler must predict the probability of being relevance in unvisited web page. To train classifier that prioritizes the frontiers reinforcement learning, context graph, text of linking pages is used which in turn guides the crawler. The performance of a focussed crawler is measured based on the richness of links retrieved corresponding to the specific topic search. Seed selection significantly influences the efficiency of a crawler. A white listing strategy is used to prevent running of unauthorized programs thus focussing the crawl from a list of high quality seed URLs and also limits the scope of crawling to these URLs domain. These white list programs should undergo periodic updating.

2. CATEGORIZATION OF FOCUSED CRAWLER
The categorization of the Focused crawlers can be done as follows (Jaytrilok Choudhary, Devshri Roy, 2013) [4]:

(i) **Classic Focused Crawlers**, where the input is a user query describing the topic, initial seed. Pages that are pointed by higher priority links are downloaded first. For assigning download priorities certain criteria are followed based on their likelihood to topic query related pages. The downloading priorities are also based on the similarity between the topic and anchor text of the linked pages, or it may be between the topic and text of the page containing the link, relating to the pages containing the search topic query.

(ii) **Semantic Crawlers**, where the semantic similarity is applied to determine the page to topic relevance from this the downloading priority is assigned to the pages.

(iii) **Learning Crawlers,** where a training process is applied for assigning visiting priorities to web pages. A training set is given to the learning crawler which consists of both the relevant and non relevant web pages. Those pages that are relevant to the topic are assigned with higher visiting priority.

3. PRIORITY BASED SEMANTIC WEB CRAWLER
In priority based semantic web crawling (Jaytrilok Choudhary, Devshri Roy, 2013) [4], priority queue is used as a database to keep URLs with the corresponding semantic score, which is calculated by Ontology and Vector Space Model. For unvisited URL, anchor text semantic similarity score is used. To crawl next, the URL with maximum score is returned from the queue. The similarity score is calculated using the formula

$$\text{Term Weight } w_t = tf_t * IDf_t \quad (1)$$

Where tf_t is term frequency, IDf_t is inverse document frequency, w_t is term weight

$$IDft = \log(\frac{D}{df_t}) \quad (2)$$

Where D is given to the total number of web pages lies below parent pages, df_t is given for number of pages where term appears. Then cosine similarity is calculated for each concept term in topic ontology and each word.

3.1 Working of Priority based Focused Crawler

1. It starts with initial seed, and then downloads WebPages at given seed.
2. Finds out all new URLs present in the page that is downloaded earlier.
3. Then corresponding new URLs are downloaded then the semantic similarity is calculated.
4. The semantic score thus calculated and the page is added to the Priority Queue and returns with maximum semantically scored web page.
5. Finally new URLs are again downloaded and the process will be continued again.

The efficiency of the crawler is calculated based on the harvest rate. It is given by the formula as

$$\text{Harvest Ratio} = \frac{\text{No.Of relevant web pages crawled}}{\text{Total number of web pages crawled}} \quad (3)$$

The higher harvest ratio, higher the performance. The crawling performance of priority based semantic crawler over simple crawler is 88%, over focused crawling is 28%, priority based is 6%.

Two crawling strategy is used in web crawling. They are Breadth First crawling and Best First Crawling.

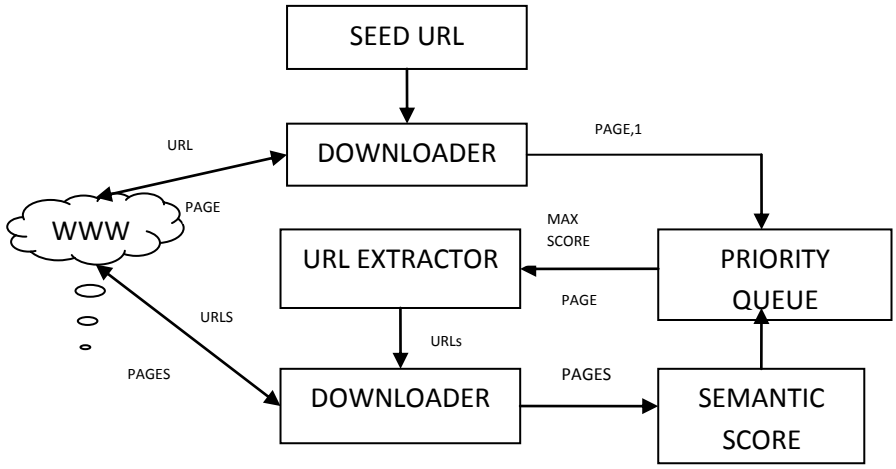

Figure 1: Architecture of Priority Based Focused Crawler

Table 1. Comparison of Breadth First Crawling and Best First Crawling

Breadth First Crawling	Best First Crawling
They are called as classical web crawler. For seed links they download web pages from initial seed.	This downloads only relevant web pages of a particular given search topic.
Until a specific count is achieved, new URLs are fetched from the download web pages.	The crawler that uses this strategy is called Focused Crawler.
Then these URLs are added to queue, and then crawling is repeated by fetching one by one URL.	It is an extension of breadth first crawling.

In Semantic Web Crawling, the semantic similarity between topic and web pages is calculated. As users, use natural language for information retrieval it is quite difficult to express the meaningful information in short term texts, hence challenging a computational problem. Thus increasing focus on computing similarity. Similarity is identified by performing aggregation on similarity values of all pairs. Another technique of word co-occurrence is pattern matching where local structural information is used from predicated sentence. Thus, limited set patterns convey a meaning thereby providing generalisation (Stephen J. Green, 1999) [8].

Semantic similarity is computed by comparing the semantic vector from the corpus. Similarity methods can be categorised as edge counting based and information theory. An automatic method to estimate the semantic similarity between words or entities can be done using page counts and snippets. A snippet can be used efficiently for top ranking query result. So, there is no guarantee that similarity can be measured from top ranking snippets. Indicative phrases are used for finding the similarity (Stephen J. Green, 1999) [8]. Another way of finding semantic similarity is using WordNet, where three methods can be employed like Node Based Method, Edge Based Method and Hybrid Methods. (a)Node Based Method is used to estimate the semantic similarity from related words in WordNet by computing the amount of information. (b) Edge Based Method calculates the distance of edges covered on the shortest path between words in WordNet. (c) Hybrid Methods combines the information content theory and structure information from WordNet, hence estimating the semantic similarity (Danushka Bollegala,2011) [1]. Apart from several other methods topic ontology is also used to find the semantic similarity between topic and web page (Jaytrilok Choudhary, Devshri Roy, 2013) [4].

Table 2. Comparison of Latent Semantic Analysis and Hyperspace Analogues to Languages

Latent Semantic Analysis(LSA)	Hyperspace Analogues to Languages(HAL)
A complete model for understanding a language.	It builds a word by word matrix.
A set of representative words are identified from a large number of contexts.	This matrix doesn't catch the sentence meaning well.
A context matrix is formed containing presence of words.	Diluted Sentence Vector.
Dimensionality is reduced.	Euclidean distance is used.
It is appropriate for larger text.	This Distance measurement gives the sentence similarity.

4. FOCUSED CRAWLER BASED ON ONTOLOGY

The web engines index has to be fully qualitative for successful result in relevance of search. In order to serve particular user needs and to reduce the effort of developers of using large volumes of data, vertical search engines are implemented. The important problem that has been addressed in implementing focussed crawler is to make a care in the effectiveness of identifying topic-relevant web pages and also exploits the facilitation of crawler's decision making policy. The crawler usually downloads resource that comes across, irrespective of its content, quality of usefulness to search engine. So the proposed approach in (Lefteris Kozanidis, 2008) [5] is to leverage the number of resources and to tools to compile rich knowledge base, from which a crawler can make decision to download pages of high priority. There are two complement programs to design the topical focused crawler. They are:

(i) To detect the topical content of a web page, a classifier is integrated with topical ontology, thus computing the relevance of the page to the topic.

(ii) To extract the text nugget that is closest to the pages topical content, passage extraction algorithm is used.

During web walkthrough, a crawler checks the knowledge base of training examples to decide whether a new page is to be downloaded. For this the knowledge base is built with topic relevance values and topic similarity extracts of more number of pages. In the proposed work of (Lefteris Kozanidis,2008) [5], the thematic terms referring sequences of semantically related term are generated by lexical chaining approach, exploring the WordNet lexical ontology.

After extracting keywords from page's content, topical ontology is used and these keywords are related to corresponding nodes in ontology topic [thematic content] computation.TODE topical ontology has been used here for web page's classification. Based on the matching of thematic keyword of pages, the degree of every ontology topic is calculated, which refers to the page's content as the fraction of thematic terms matching the topic T. Probability of pages belonging to Ontology topic is estimated by

$$\text{Topic Relevance (P)} = \frac{|\text{Thematic Keywords in P matching T}|}{|\text{Thematic Keywords in P}|} \quad (4)$$

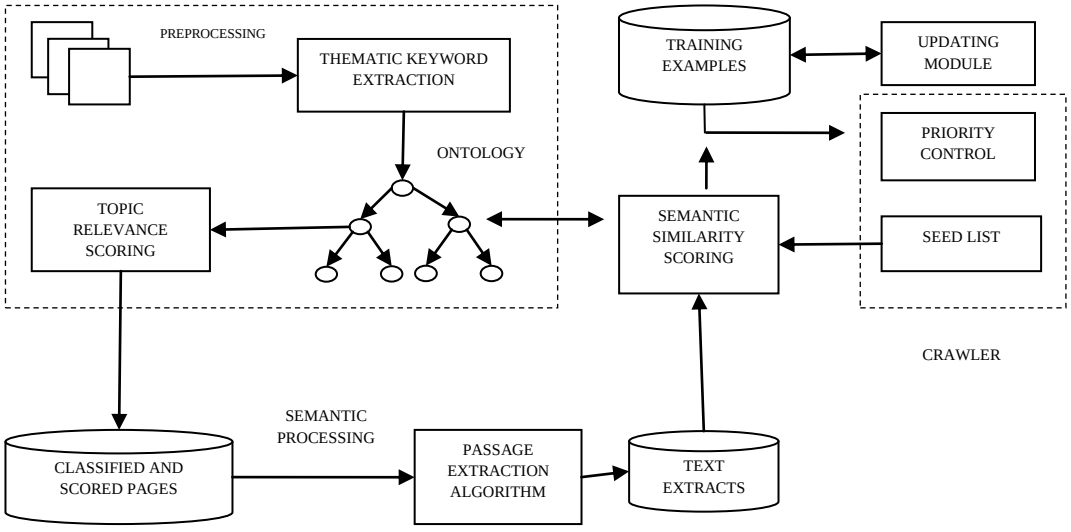

Figure 2: Focused Crawler Architecture with Ontology Technique

5. CELL LIKE MEMBRANE COMPUTING IN FOCUSED CRAWLER

In the paper (WenJun Liu, YaJun Du n,2014) [9], the problem literature is the weight factors given to the unvisited hyperlinks with topical priorities. So to solve the problem cell like membrane computing, an optimization algorithm is used in novel focussed crawler. A directed graph where pages are nodes and hyperlinks are edges. Web crawler's traverse pages in breadth first search algorithm till all pages are collected or there is no vacant storage space. This leads to the focussed crawler, were only topic relevant web pages are gathered. This reduces the massive time and space resources and the user satisfaction is attained. The two phases in crawling process are: determination of initial

URLs seed, selection of better unvisited URLs. Initial seed includes topical seeds and generic seeds.

5.1 Topical Seed

These are selected from the retrieved results, by inputting the topic-relevant keywords.

5.2 Generic Seed

These are hyperlinks pointing to top list web directory pages in the hierarchy.

All unvisited URLs are extracted from crawled pages and variants texts of each unvisited hyperlinks are acquired. Similarity between these texts and the topic are measured using information retrieval model like vector space model.

Priority of each unvisited hyperlink is obtained by integrating these topic relevant similarities of different texts of hyperlinks. Priorities of hyperlinks were used to detect the traversing order of these unvisited URLs. Focussed crawlers have to determine the document types of unvisited URLs. Semantic similarity retrieve model was put forward to make focussed crawlers retrieve pages with semantically similar terms. It also computes priorities of unvisited hyperlinks. Priority of a hyperlink was given by structural fragments of pages. These fragments are hyperlink, heading of sections, surroundings of paragraphs, even table captions and image descriptions containing the hyperlink. Two approaches are used for computing topic similarities exists. a) VSM approach b) SSRM approach.

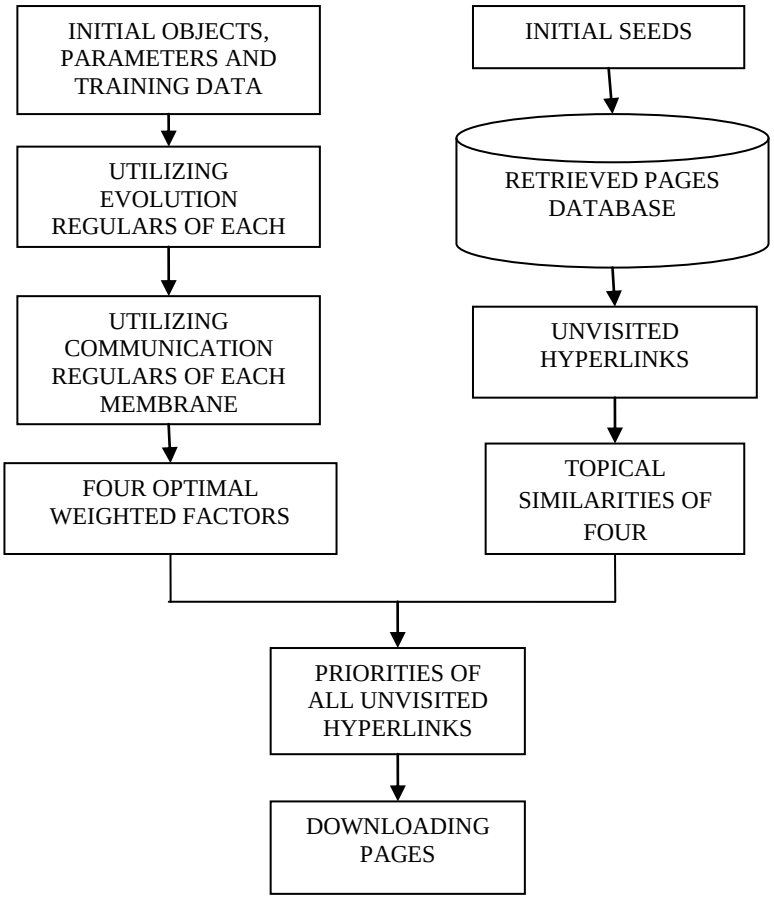

Figure 3: Novel Focused Crawler with Cell Membrane

In the cell like membrane computing focused crawler, there will be four documents corresponding to the unvisited hyperlinks. The four documents include

- Full text of pages
- Anchor texts
- Title texts of pages
- Surrounding texts of paragraphs

For each membrane, an object is initialized by random generation and some parameters are set manually. For each object, four weighted factors are assigned, which provides the degrees of similarities between these four documents. For the optimal weighted factors, the CMCFC adopts the evolution regulars and communication regulars off all membranes, this avoids the falling of CMCFC into local optimal solution. The evolution regulars include selection regulars, crossover regulars, and mutation regulars. They are adopted until the generation is achieved. From this the optimal four weighted factor is outputted by CMCFC, which has the maximum fitness values and Root Measure Square error of hyperlink priority is achieved with minimum values. To download initial web pages that have to be stored in database, initial seeds are used by the crawler. From this hyperlinks are extracted and then it is added in an unvisited list.

5.3 VSM Crawler

It uses cosine similarity to find the priorities. The unvisited hyperlink that is considered here is a full texts and anchor texts of pages. Topical similarities of documents of unvisited

hyperlinks are calculated by inner products between document and topic vectors. By linear integration of topical similarities of full texts and anchor texts the priority of each unvisited hyperlink is computed. The term vectors are acquired by the term frequency inverse document frequency and the cosine similarity between two terms. Optimal weighted factors are provided by membrane computing method. The degrees of similarities of four documents are given by weighted factors, of each unvisited hyperlinks. The documents of unvisited hyperlink include the full texts, anchor texts of pages (WenJun Liu, YaJun Du n, 2014) [9]. The standard formula for the computation is

$$\text{sim(d,t)}=\vec{d}.\vec{t}\frac{\sum_{i=1}^{n} w_{di}\, w_{dt}}{\sqrt{\sum_{i=1}^{n} w di^{2}}\sqrt{\sum_{j=1}^{n} w^{2}_{tn}}} \qquad (5)$$

Where, d is the document, vector refers to term vector of the document and topic respectively.

5.4 SSRM Crawler

It uses hyperlinks to compute priorities. Topical similarities of documents are sorted out by associating term frequencies and term semantic similarities and accumulating these products. The Semantic similarities between two terms depend on the simulation between concepts corresponding to two terms. Semantic similarity between terms is computed with the methods of edge count for information content (WenJun Liu, YaJun Du n, 2014) [9].

6. CELL LIKE MEMBRANE COMPUTING IN OPTIMIZATION

A framework for devising computing models such as cell-like, tissue-like and neural-like models. The computing devising obtained from Membrane computing is a distributed parallel model, where mutisets of objects are processed based on the rules. There are three parts in P systems. They are: membrane structure delimiting the mutisets, mutisets of objects, that are evolved based on the rules. In cell-like membrane computing there will be a hierarchical arrangement of membranes where object mutisets are placed, rules for evolving the objects is given, such as multiset rewriting rules similar to common chemical reactions. The hierarchy of membranes can also be changed based on addition and removal of membranes by cell division methods. For instance, strings are the objects, evolves string processing rules such as insertion, deletion (Gheorghe Paˇun, 2010) [2].The optimal weighted factors are provided by membrane computing method. The degrees of similarities of four documents are given by four weighted factors of each unvisited hyperlinks. The calculation model of membrane computing, called P system is non deterministic and is a parallel model with hierarchical structure. In membrane computing each membrane of P system is taken as an individual calculation unit for specific calculation. The non determinism of P System involves the selection of the objects to regulate the computation thus completing the task in parallelism. The probability of an object getting selected is based on fitness value of the object.

7. SELF ADAPTIVE SEMANTIC FOCUSED CRAWLER

Self adaptive semantic focused crawler [SASF] framework (Hai Dong ,2014) [3], used to discover format and index mining service information by considering for the three issues by using ontology learning for maintain the performance of the crawler. The new concept involved in the SASF crawler are Vocabulary based ontology learning and hybrid algorithm for matching semantically relevant concepts and metadata.Discovering the service or service information in particular environment is done automatically or semi-automatically. SASF framework uses Semantic Focused Crawling to solve the above the problem of heterogeneity, ubiquity, ambiguity in service discovery. Heterogeneity refers to the problem of classification of service advertisement. Ambiguity refers to the problem of identifying the service information that doesn't have a consistent format and standard. Ubiquity refers to the problem of discovering the registries of services that are geographically distributed. SASF crawler is used for helping search engines to precise the mining of service information. Supervised ontology learning method is used to maintain the harvest rate of the crawler.The input given to the supervised learning method is the domain and the topic represented by a concept. It may work with in an uncontrolled web environment. Unsupervised ontology learning method where the input is topic and relevance score of the topic. Metadata crawler contains information such as content, length and length variation, value based analyses, frequencies, patterns, domains, dependencies, relationships. Determines the semantic relatedness between concepts and metadata concept metadata using semantic similarity algorithm . In this algorithm the semantic similarity between concept description and service description is measured. It follows a hybrid pattern by aggregating two algorithms namely semantic based string matching algorithm and statistics based string matching algorithm.

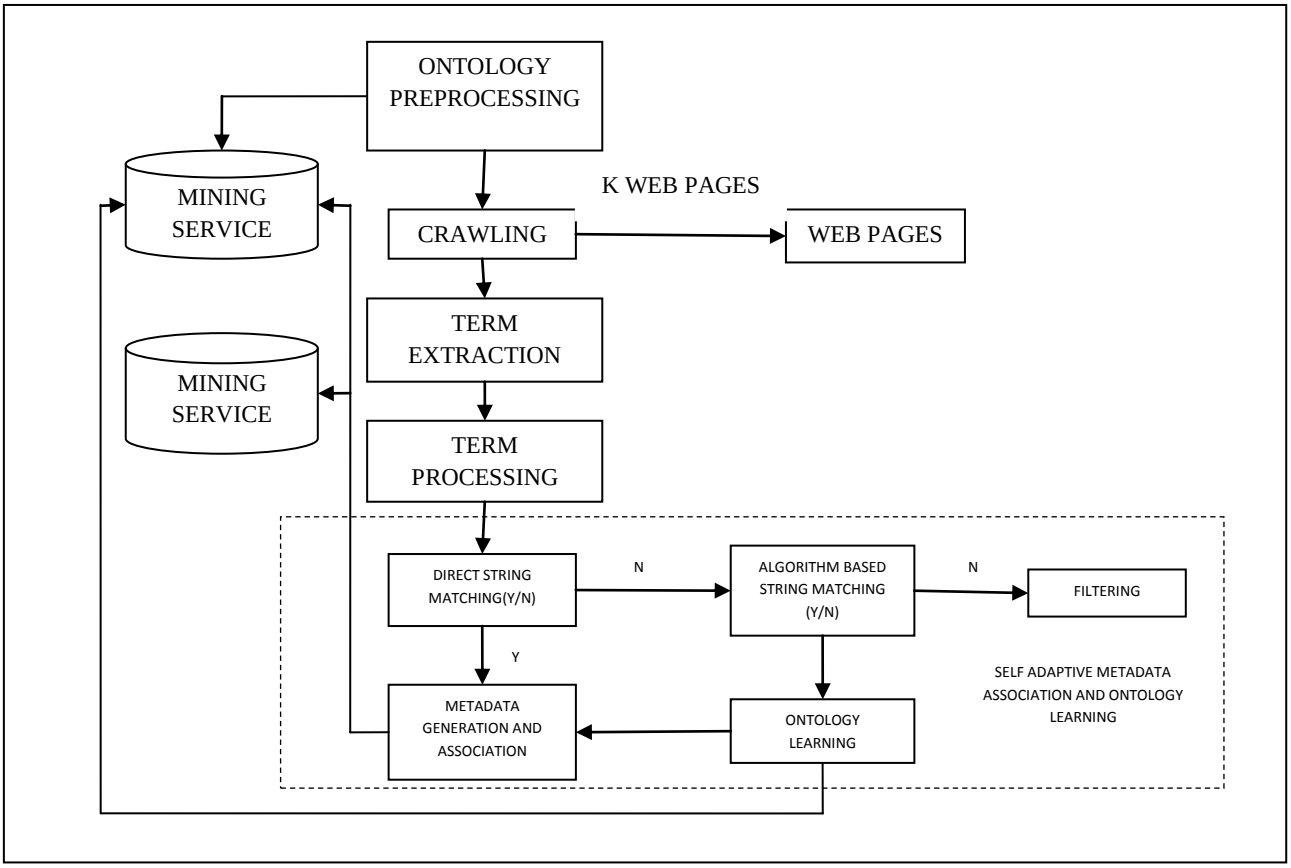

Figure 4 : Architecture of SASF Crawler

Table 3. Comparison of SeSM and StSM

SEMANTIC BASED STRING MATCHING ALGORITHM (SeSM)	STATISTICS BASED STRING MATCHING ALGORITHM (StSM)
It measures the text similarity between concept description and service description, which are considered as two groups, using the semantic similarity model and WordNet	It is a complementary method of SeSM, as it may not work effectively under certain circumstances.
Resnik information theoretic model is used to find the semantic similarity between those two groups.	Example: Service Description-"Old mining workings consolidation Controls" Concept Description-"mining contractor". The similarity value is (1+1)/5=0.4
Plebani's bipartite graph model is used to assign the matching between terms among the group in an optimised way.	The actual relevance value is relatively higher in StSM algorithm
In WordNet, the relative position of two concepts is compared to measure the semantic similarity	In StSM, an unsupervised training method, is used to find the maximum probability that a concept description and service description co-occurs in webpages.

Table 4. Comparison of harvest rate of five crawlers

S.NO	METHODS	HARVEST RATE
1	Breadth First Crawler	0.546
2	SSRM Crawler	0.604
3	VSM Crawler	0.661
4	CMCFC	0.763
5	SASF Crawler	0.6

8. CONCLUSION

Crawlers are designed to download web pages from internet. There are several crawlers has been designed for specific functionalities. Among them two crawlers, SASF crawler and CMCFC crawler has been taken for comparative study. Self Adaptive Semantic Focused Crawler was designed for formatting and indexing mining service information from the internet. The harvest rate of SASF crawler is 0.6. A cell like membrane computing that has been used in crawler for optimization improves the performance of the focused crawler. The harvest rate of CMCFC is 0.763. In Self Adaptive Semantic Focused Crawler it follows unsupervised learning framework in ontology learning and a concept-metadata matching algorithm is used for finding relevance between service concept and service metadata. A Semi Supervised framework can be used in future work to improve the performance of the crawler by estimating a threshold value which sets boundaries for concept matching.

9. REFERENCES

[1] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, Member, IEEE (2011),"A Web Search Engine-Based Approach to Measure Semantic Similarity between Words". IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 7.

[2] Gheorghe Pa˘un, (2010.)"A quick introduction to membrane computing", The Journal of Logic and Algebraic Programming Vol. 79 , Pages 291–294, August 2010.

[3] Hai Dong, Member, IEEE, and Farookh Khadeer Hussain,(2014)" Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery", IEEE Transactions on Industrial Informatics, vol. 10, no. 2.

[4] Jaytrilok Choudhary, Devshri Roy,(2013),"Priority based Semantic Web Crawler ", International Journal of Computer Applications ,Vol. 81, No 15.

[5] Lefteris Kozanidis,(2008), "An Ontology-Based Focused Crawler ",NLDB '08 Proceedings of the 13th international conference on Natural Language and Information Systems: Applications of Natural Language to Information Systems, Pages 376 – 379,Springer-Verlag Berlin, Heidelberg ©2008.

[6] Nidhi Jain, Paramjeet Rawat,(2013), "A Study of Focused Web Crawlers for Semantic Web", International Journal of Computer Science and Information Technologies, Vol. 4, No 2,Pages 398-402.

[7] Sotiris Batsakis, Euripides G.M. Petrakis, Evangelos Milios,(2009) ,"Improving the Performance of Focused Web Crawlers ",Data & Knowledge Engineering archive Vol. 68,No 10, Pages 1001-1013.

[8] Stephen J. Green,(1999), "Building Hypertext Links by Computing Semantic Similarity", IEEE Transactions on Knowledge and Data Engineering, Vol. 11, No. 5.

[9] WenJun Liu, YaJun Du n,(2014), "A novel focused crawler based on cell-like membrane computing optimization algorithm", School of Mathematics and Computer Engineering, Xihua University, Chengdu 610039, NeuroComputing, Journal Homepage of Elsevier China.