4. **Focused Crawlers for Web Content Retrieval**

The World Wide Web is a huge collection of web pages where every second, new piece of information is added. Finding relevant web resources indeed is a protracted task and searching required content without any explicit or implicit knowledge adds more intricacy to the process. Generic crawlers traverse complete web in order to generate indexes which are used later for searching and recommending links to users. This method leads to huge storage space requirements and usually falls short to cope up with the dynamic nature of the Web. Focused crawling in such scenarios provides a better alternate to generic crawling especially when topic specific and personalized information is required.

This chapter proposes and discusses the designs of different types of Focused Crawlers that collects potentially relevant learning content from different types of web repositories. Prototypes of two such Focused Crawlers, FCHC (Focused Crawler based on Human Cognition) which explores Social Bookmarking Site for the useful content and, DSRbasedSFC (Dynamic Semantic Relevance based Semantic Focused Crawler) which crawls on the WWW, have been designed and discussed.

4.1 **Introduction**

The search engines in general and web crawlers in particular are facing challenges of ever increasing volume of the WWW. Every day thousands of web pages are being added to the web. With the passage of time, it is becoming difficult to crawl and update the complete web in short time. In such circumstances a goal-directed crawling or Focused crawling is a promising alternate solution to a generic crawler. A focused crawler, the term coined by Chakrabarti et al. (Chakrabarti, Berg, & Dom, 1999) is a topic-driven web crawler which selectively retrieve web pages that are relevant to a pre-defined set of topics. A focused crawler yields latest resources (web pages) relevant to the need of individuals while utilizing minimum storage space, time and network bandwidth (Batsakis, Petrakis, & Milios, 2009). Applications of the focused crawler include business intelligence (to keep track of publically
available information about their potential competitors) (Pant, Srinivasan, & Menczer, 2004), generating web based recommendations, retrieving domain/topic relevant e-Learning web resources scientific paper repositories (Zhuang, 2005), (Moghaddam, 2008) and many more. They are also useful to update topic relevant indexes and web portals where specific information is required to fulfill the community’s information need, in comparatively much lesser time. Dong and Hussain (2011) have shown their use in industrial Digital Ecosystems for automatic service discovery, annotation and classification of information. In e-Learning the crawlers can be trained to collect learning content related to a specific topic for a learner as shown in this chapter.

The focused web crawlers are designed to retrieve web pages based on various approaches or criteria that identify relevant pages or/ and priority criterions to sequence the web pages to be crawled and add them to the local database. This database may then serve different application needs. Focused crawlers are grouped into two broad categories (Figure 4-1) namely, the Classic Focused Crawler and the Learning Focused Crawlers (Batsakis, Petrakis, & Milios, 2009). Both have their own variations depending on various algorithms (Pant, Srinivasan, & Menczer, 2004) applied on them. However the main difference between the two is that the former follows the predefined and fixed guidelines or criteria for crawling whereas the latter learns or adapts the crawling guidelines based on the dynamically updating training set. Learning focused crawlers need comparatively more preprocessing time for building and updating the training set, which usually involves additional generic crawl to fetch web pages, their manual segregation into good and bad web pages for every topic and then applying some learning technique to determine the relevance score of

![Focused Web Crawlers Taxonomy](image_url)

**Figure 4-1: Focused Web Crawlers Taxonomy**
web pages (Zheng, Kang, & Kim, 2008). Semantic crawler and social semantic crawler are the variations of the Classic Focused Crawler which determines the web page relevance by utilizing a preexisting knowledge base. However, they may also be extended to Learning Focused Crawlers at the expense of preprocessing time.

4.2 Related Work

The overall performance of a focused crawler mainly depends on the method of determining the priority of web pages to be crawled, which affects the harvest ratio\(^{32}\) of a focused crawler. The priority computation usually includes methods to determine the relevance of web pages, and/or the path to reach relevant web pages. Therefore the major task during the focused crawl is to predict the ordering of web page visits. Some early designs of Focused Crawlers parsed anchor text to compute the relevance of web pages (Craswell, Hawking, & Robertson, 2001). The web page relevance was also predicted by analysing the link structure and content similarity (Jamali, Sayyadi, Hariri, & Abolhassani, 2006).

A similar research (Hati & Kumar, 2010) calculated the link score based on average relevance score of parent web pages and division score (keywords related to the topic category) to determine the web page relevance. This was computed by taking term frequency of top ten weighted common words from a set of seed pages which in turn was a set of common URLs retrieved using three search engines. Such approach, in some particular cases may yield URLs from an undesired domain, which consecutively may result in wrong fetches. However extracting search topic related keywords from a domain ontology eliminates the problem of selecting out-of-context keywords. Moreover computing Semantic relevance instead of hyperlinked structures or PageRank algorithms (Page, S. Brin, Motwani, & Winograd, 1998) overcomes the problem of Search Engine Optimization (SEO basics, 2008), (Google, 2010), (Callen, 2007).

There exist a few publications on focused crawlers that utilizes ontology for varied purposes, viz. Luong, Gauch, and Wang (2009) has used amphibian morphology ontology to retrieve web documents and the Support Vector Machine

\(^{32}\) fraction of relevant web pages among total crawled web pages
focused crawlers need extra crawling efforts to build either a training-set or generating prerequisite data. Moreover, most of the existing focused crawlers are computationally expensive.

The use of collaborative social tagging for gathering relevant web resources is comparatively a new area of research in Information Retrieval, especially educational content retrieval, with only few published works. Collaborative web page tagging is one of the powerful means to build a folksonomy which may be consumed as implicit feedback for determining web resource relevance. However, investigations on collaborative tagging systems, kind of tags, their distribution, suitability and usage for improving search have been done extensively. Bischoff et al. (Bischoff, Firan, Nejdl, & Paiu, 2008) have showed that most of the tags in a collaborative site can be used for search, and in most cases tagging behavior exhibits approximately the same characteristics as the searching behavior.

In another study by Valentin et al. (Valentin, Halpin, & Shepherd, 2009), it was shown that the tagging distribution of heavily tagged resources tends to stabilize into power law distribution. This implies that the information driven by the tagging behavior provides a collective consensus around the categorization.

A Social Semantic Focused Crawler can utilize the Social Web and semantic knowledge to gather relevant web resources. Web 2.0 which is considered as a Social Web comprises of various blogs, Social Bookmarking Sites (SBS), facebook, twitter, flicker, etc. where web users are allowed to share and organize their information and add their objects (text, images, videos etc.) to the sites to represent their views. A single point of access to various social network systems (Chao, Guo, & Zhou, 2012) would give more benefits, as the search area and number of users would get
assimilated. However, at present the crawlers need to be designed for every site individually.

Though much of the published literature has shown social data, in particular, SBS as a vision and a promising solution for better search results, few have also raised some concerns over its limitations and complexities. For example, Greg et al (Pass, Chowdhury, & Torgeson, 2006) noted an increase in noise while mapping tags and documents over a period of time. In fact a huge data provided by SBS needs a proper investigation. A few researchers (Chen & Yi, 2009), (Bao, et al., 2007) have also suggested their view points in this regard. According to them, effective methods can be developed to re-rank the search results using the tagging information from SBS.

Zanardi & Capra (Zanardi & Capra, 2008) have used the user similarity between the querying users and the users who have already created tags using the topic terms in SBS, to rank the relevant resources. The word ‘query’ in their system resembles the search topic used in our approach. They used a two step model to find relevant tagged web resources from a SBS by finding user similarity. The first step was expanding the query term based on query tags chosen from a folksonomy. The second step ranks the SBS resources by finding similarity between socially annotated tags. Wu et al. (Wu, Zhang, & Yu, 2006) derived semantics statistically from social annotations. They used probabilistic generative model to obtain semantics by analyzing occurrences of web resources, tags and users.

In another interesting work on Upper Tag Ontology (UTO) (Ding, et al., 2010), the information on various SBS are restructured into ontology which can be queried to determine varied relationships among users, tags and resources. However, the noise in social sites is one of the biggest constraints in building the knowledge base. The proposed design of the Focused Crawler in this paper can efficiently be used to filter out this noise by downloading only those web resources that are likely to be relevant.

The work by Bao et al. (2007) uses social similarity ranking and social page rank to rank relevant resources. The analysis is based on web page annotations and web page annotators’ profile. However, they used keyword similarity method to
associate query with annotations, where as relevance ranks have been computed by analyzing web pages and social annotations solely. Our approach, instead, analyzes social annotations and the Semantic Relevance which is computed through the Concept Ontology (as described in Chapter 3).

Existing work related to Semantic Focused Crawlers mainly focuses on building them using ontology, but the way ontology is being utilized by these crawlers depends on the search motive. Thus, Semantic Focused Crawlers can be categorized, based on the search motive, into two types. One of them is specifically designed to search relevant ontologies in the WWW and the Semantic Web, usually used for ‘ontology search engines’ such as Swoogle (Ding, et al., 2004), OntoKhoj (Patel, Supekar, Lee, & Park, 2003), OntoMetric (Lozano-Tello & Gómez-Pérez, 2004), or AkTiveRank (Alani, Brewster, & Shadbolt, 2006). They crawl ontology repositories to gather linked data (in rdf, xml or owl format) existing on the Web. Hence at the core level, they search ontologies and rank them according to the concept density within ontology. The other type of the Semantic Focused Crawlers search relevant web pages (documents and not ontologies) from the Web by utilizing a pre-existing semantic knowledge to determine web page relevance. Thus, the former retrieves relevant ontologies while the latter retrieves semantically relevant web pages. Our proposed approach focuses on the latter type of Semantic Focused Crawlers. They are sometimes also referred to as Ontology based Focused Crawlers.

The literature has few reviews on Ontology based focused crawlers or Semantic Focused Crawlers (Dong, Hussain, & Chang, 2008), (Dong, Hussain, & Chang, 2009). Ehrig et al. (2003) compute relevance score by establishing entity reference using tf/idf weights on natural language keywords and background knowledge compilation based on ontology. They refer ontology at each step to gather the relevance score which may make the computational time expensive. Moreover tf/idf algorithm is usually applied on a large corpus to use it effectively (Garcia, 2006b), (Salton & Buckley, 1987).

The approach by Diligenti, et al. (2000) uses context graphs to find short paths that leads to relevant web pages. THESUS crawler (Halkidi, Nguyen, Varlamis, & Vazirgiannis, 2003) organizes web page collection based on incoming links of a
web page and thereby cluster the web pages.

The Ontology based Web Crawler proposed by Ganesh et al. (2004) computes similarity between web pages and ontological concept by exploring association between parent page and children pages whereas courseware watchdog crawler (Tane, Schmitz, & Stumme, 2004) is built on the KAON system (Maedche & Staab, 2004) which utilizes the user feedback to the retrieved web pages. However, both these papers have not discussed the evaluation details of their conceptual framework. LSCrawler (Yuvarani, Iyengar, & Kannan, 2006), a general focused crawler, built to index web pages by computing similarity between web pages and a given topic shows better results comparisons with a full text crawler.

The proposed Focused Crawlers, FCHC and DSRbasedSFC explores Social Bookmarking Site and the WWW respectively, to gather relevant web resources. Unlike other Social Semantic Focused Crawlers FCHC computes weights of tagged resources using the Concept Ontology and while crawling collects the ‘relevance’ information as well. The DSRbasedSFC computes the Semantic Relevance of web resources dynamically during the crawling, which again makes use of the Concept Ontology.

4.3 Focused Crawlers

Web crawlers, also known as bots, spider etc. are software programs that are designed to assemble URLs and other web page attributes locally by exploring the link structure of the web. Focused Crawlers are the specialized form of the web crawlers that selectively seek out web pages that are relevant to a topic. The baseline idea of the focused crawler is to maximize the retrieval percentage of relevant web pages while keeping the total number of fetched pages at the minimum (Chakrabarti, Berg, & Dom, 1999). Therefore for a focused crawler it is always crucial to search for relevant group/cluster of pages on the web. There exist many algorithms that predict the probability of getting relevant clusters through different link paths (Pant, Srinivasan, & Menczer, 2004). A more detailed description of the two categories of focused crawlers (Figure 4-1), Classic focused crawlers, their variations and Learning focused crawlers is given in the following sub-sections.
4.3.1 Classic focused crawler

Figure 4-2 shows the basic flow of a Focused Crawler. The simple Classic Focused Crawler, Social Semantic Focused Crawler and Semantic Focused Crawler are all its variants which differ by the criteria they apply in selecting the crawling area, determining the type of priority function or method of determining the page relevance. The basic crawler uses a set of pre-selected seed URLs to initiates the crawl. The crawler first fetches a page and then parses it to extract all the links and to check topic relevance of the page, based on the page relevance criterion.

The page relevance criterion is similar to the classifier (Chakrabarti, Berg, & Dom, 1999), which evaluates the relevance of hypertext document. All the parsed URLs are enqueued in the priority queue if the page is considered relevant according to the page relevance criterion. The priority of the page is further governed by the page priority criterion, which resembles the distiller (Chakrabarti, Berg, & Dom, 1999).

![Figure 4-2: Baseline Flow of Classic Focused Crawler](image-url)
Retrieving and Organizing Web Resources Semantically for Informal E-Mentoring 1999). The distiller identifies hypertext nodes that have good access points to many relevant pages within few links. Different priority queues may be used here depending on the application’s requirements. If a page is considered non-relevant then the links on that page are not inserted to the queue and hence that page is not crawled further. This is where it differs from a generic crawler which is used by search engines. After checking the termination condition, which could be the number of URLs to be crawled, time limit or the empty queue, the crawler decides to either stop or dequeue a top priority URL from the queue to repeat the crawl process. There exist many variations to this basic focused crawler by varying the four important selection criterions. i) The seed selection criterion, ii) Inject clusters to be crawled, iii) Page relevance criterion and iv) Page priority criterion.

One such prominent variation of the classic focused crawler is Semantic focused crawler that uses semantic knowledge (hyponyms, hypernyms, synonyms,

![Diagram of Learning Focused Crawler](image.png)

**Figure 4-3: Learning Focused Crawler**
antonyms etc. or the topic related words) to define page relevance criterion and variation to this is Social semantic focused crawler (the proposed crawler) which also specify an area on the web to be crawled. In our case the area of the web is Social Bookmarking Site. These are explained in subsequent sections.

4.3.2 Learning focused crawler

The other type of category of the focused crawlers is Learning Crawler (Figure 4-3). These crawlers apply a training set to govern one or more criterions of the classic focused crawler. The training set consists of a set of example pages related to the topic. The crawler identifies relevant and non-relevant pages by making use of the training set. Learning crawlers use various methods based on Bayesian classifier, context graphs (Diligenti, Coetzee, Lawrence, Giles, & Gori, 2000) and Hidden Markov Model (Batsakis, Petrakis, & Milios, 2009) to estimate the link distance between a crawled page and the relevant pages. This data is used to train the crawlers by setting the criterion for page relevance and fetching priority.

4.4 The Proposed Approaches for Crawling

This section presents designs of two types of Focused Crawlers one each from the variations of Classic Focused Crawlers, i.e., Semantic and Social Semantic. They crawl on different types of Web areas using different approaches. The first crawling approach called Focused Crawling based on Human Cognition (FCHC) collects web resources from a collaboratively tagged bookmarks (folksonomy) site those are likely to be semantically relevant to a given topic. The approach has been given using two different search patterns, Breadth-First Pattern (BFP) and Depth-First Pattern (DFP). A mechanism to compute Social Semantic Relevance of web pages with respect to a given topic is also proposed for FCHC approach. It uses a list of semantically expanded terms (explained in Chapter 3) on a given topic which is obtained using the domain ontology. The list also includes the Semantic Relevance for each expanded term. These expanded topic terms are matched with the tagged bookmarks during SBS crawl using the FCHC search pattern which reduces the sparse data from getting downloaded. The FCHC approach reflects on both of the important factors, i.e. the Semantic Relevance of tags with the search topic and the popularity of web resources among the community.
The other proposed Focused Crawler called *DSR* based Semantic Focused Crawler uses the Concept Ontology (Chapter 3). A unique feature of this crawler is that it uses Dynamic Semantic Relevance (*DSR*) to prioritize the crawling list of the fetched web pages. The weights used to determine the semantic distance between two concepts in the proposed crawler are computed from the domain ontology which as per our knowledge, are assigned manually in the work done in literature and stored in ontology for semantic computation.

### 4.4.1 Focused Crawler using Human Cognition

The selection of seed URLs, which directs a focused crawler to identify a search area on the Web, is perhaps one of the major criteria that affect the results of a focused crawler. This feature of focused crawler motivated us to apply the focused crawl on a social site, where web users belonging to various communities with varied interests, across the globe pool up web resources of their interest and bookmark them for later referral. Recommending semantically relevant web resources from the collection of manually annotated web resources is another source of motivation. A site consisting of such a collection, called a Social Bookmarking Site (SBS) is an online social network and a collaborative bookmarking system that facilitates web users to organize bookmarks to the web resources of their interest. Social networks have been analyzed for a couple of decades to find useful information related to web pages. The web users opt for relevant keywords to tag the web resources of their interest so that they could easily refer them later without the repetitive searches on the Web. This information therefore can be used to find relevant web pages for other users as well.

#### 4.4.1.1 FCHC Crawler Framework

Social Semantic Focused Crawlers utilize the Social Web and semantic knowledge to crawl relevant resources from the select portions of the Web. In the proposed work, the search pattern used by the crawler mimics a pattern usually followed by human users while searching web pages of interest in a typical SBS. Hence the crawler is named FCHC which stands for Focused Crawling based on Human Cognition.
Besides, the crawler utilizes social and semantic information to retrieve tagged web pages, and therefore belongs to the Social Semantic Crawler category.

The searching pattern used to search relevant resources in SBS makes this crawler different from others. The framework for the FCHC Crawler is illustrated in Figure 4-4. Unlike other focused crawlers which parse every web page to calculate or predict the relevance of a web page or path to reach relevant resources, the proposed crawler makes use of the tags and the bookmarks tagged by the SBS users. The crawler makes use of the Concept Ontology after initiating the search with seed URLs taken from the search results of a web search engine. The output generated by the crawler is a collection of potentially relevant bookmarks which are stored in a database for post processing. The flow graph of FCHC crawler is presented in Figure 4-5. It is apparent from the figure that the crawler follows the basic flow of crawling; however, the crawling approach used to set the criteria makes it different and possibly better from other crawlers. These important selection criteria for FCHC crawler are detailed below.
Seed selection criterion: Seed URLs are picked from the search results of a web search engine by querying on a search topic. Top N URLs are considered as seeds to initiate the crawl on the Web’s select portion. A good selection of the topic relevant URLs and the number of seed URLs, without any doubt effect the results of a crawler (Chakrabarti, Berg, & Dom, 1999). A large number of seed URLs enable a crawler to spawn a wider surface on the Web. However, in a controlled area like SBS, where the web resources are accessible through many linked points like users, tags and URLs, it becomes feasible to reach more number of resources even with a small set of seed URLs.

Provide the area for crawling: Any bookmarking site or a portal where the resources have been tagged so as they represent their inside content, can be used as an area for crawling. However, this requires designing the crawling pattern according to
the structure of the portal. In the FCHC design, a social bookmarking site delicious.com has been used for crawling purpose.

**Parse the page:** The SBS pages are parsed for extracting bookmarks which are analyzed as relevant by the crawler. During parsing, the information related to resource and tags is extracted and stored in the local database.

**Page relevance criterion:** During the crawl, the crawler checks the relevance of a page by matching the tags of the resource with the expanded topic terms. Later, after the complete crawl empirical social semantic relevance of each resource is computed using the Vector Space Model.

**Page priority criterion:** A systematic search pattern motivated by human cognition is used as the priority criterion for the crawler. Different search patterns used by the crawler are explained in the next sub-section.

**Termination criterion:** If number of URLs to be crawled is specified then the crawler uses it as a termination criterion otherwise when the priority queue becomes empty, the crawler stops.

### 4.4.1.2 FCHC Searching Patterns

The pages can be crawled using two types of pattern based on the SBS structure: Breadth first pattern (BFP) and Depth first pattern (DFP). The structure of a typical SBS and crawling patterns are illustrated in Figure 4-6. In BFP the crawler enqueue all those users who had tagged the pre-select URLs (seed URLs). Then from the queue one by one the ‘tags page’ (a page on SBS consisting of all tags marked by a user) of each user is parsed to reach resources of their interest. All these potentially relevant web resources are then added to the queue to be parsed further.

It has been noticed that a crawler, in general is hardly designed using the DFP approach, but in particular case like crawling in SBS (a controlled area of the Web), DFP has also shown promising results as we will see in chapter 6. In this pattern, the crawler first reaches the relevant resources of the first user using first seed URL and enqueue all of them. It then iterates the process for all users of the same seed URL. It similarly works upon all other seed URLs one by one.
The difference between BFP and DFP is that the latter completes retrieval of relevant tagged resources of each user one by one, whereas the former first queue up all related users and then moves to their tagged pages one by one. The social semantic focused crawler, FCHC also uses DFP to crawl on to different levels of the SBS as shown in Figure 4-6. One Level represents single iteration of resource retrieval by looking up the users and tags of the existing list of resources (for the first level these resources are the seed URLs). For subsequent levels, the crawler iterates the process using the list of resources retrieved from the previous level.

### 4.4.2 DSR based Semantic Focused Crawler

The proposed Dynamic Semantic Relevance based Semantic Focused Crawler (DSRbasedSFC) or simply SFC is a focused crawler that uses multithreading to crawl select sections of the Web which contains topic relevant web pages. The SFC utilizes the domain ontology to expand a topic. These domain ontologies are specifically designed for educational purpose to include maximum concepts that fall under a given domain in a structured way (as described in Chapter 3).
4.4.2.1 DSRbasedSFC Crawler Framework

The SFC framework is illustrated in Figure 4-7. It consists of domain ontology, priority queue, local database and the proposed multithreaded Semantic Focused Crawler. SFC runs multiple threads, where each thread picks up a top priority URL from the priority queue which is a web page with highest Dynamic Semantic Relevance (DSR). The threads independently parse the web page and extract all hyperlinks on that page. The extracted hyperlinks are added to the queue. These hyperlinks are fetched and parsed one by one to compute DSR to repeat the process of crawling. The priority queue thus, maintains the order of URLs to be parsed by the SFC threads. During the crawl each thread also checks for already visited URLs to avoid cycles. For this purpose a separate temporary queue is maintained which stores all visited URLs. All URLs of potentially relevant web pages fetched during the crawling process are stored in the local database, to be later consumed by other applications. Each thread of the SFC crawler carries out the crawl process in two parts, which are explained in detail below.
4.4.2.2 Design of DSRbasedSFC

Fetch and Parse a Web Page

The priority queue is initiated with the seed URLs, which can be fetched from a search engine. Dynamic Semantic Relevance (DSR) of these resources is computed and then enqueued to the priority queue. A web page with the top priority URL from the priority queue is fetched from the Web (shown as number ‘1’ in Figure 4-7). The web page source is then parsed to extract the URLs (hyperlinks) and tokenized to determine the frequency of each concept in the expanded concept list. This extracted data is then consumed by the next process to determine DSR.

Compute Dynamic Semantic Relevance

The dynamically computed semantic relevance $DSR_{p_i}$ of each web page $P_i$, is a distinguished feature of this Semantic Focused Crawler. $DSR$ is computed after a web page is parsed during the crawl process. Thereafter, the web page is placed onto the priority queue along with its computed $DSR$. In the next spanning iteration, the thread picks up a web page with highest $DSR$ from the priority queue so as to reach all those web pages which are linked to the parent web page. This is based on the assumption that a web page, which is considered highly relevant would contain hyperlinks to more relevant web pages, therefore the hyperlinks on this web page should be crawled first. In this way, the web pages that are more relevant to a search topic would get priority to be crawled first over the less relevant web pages.

Dynamic Semantic Relevance $DSR_{p_i}$ of a web page $P_i$ to a topic $t_0$ is computed through the following steps.

**Step 1:** Topic $t_0$ for focused crawling is expanded from the domain ontology $D_i$ where $t_0 \in D_i$.

The topic is expanded by including all parent nodes and a few levels of child nodes from the ontology. To avoid ontology access during the crawl, a structure

---

33 This particular case takes concepts up to 4th level in ontology, although this may vary according to the depth of content required on the topic.
comprising of each associated concept (term) and its Semantic Distance (explained in Step 2) is stored in a temporary memory. This reduces the time spent on frequent access and traversal of the ontology. The domain ontology for the purpose is created as a semantic graph, consisting of various concepts from the education and learning perspective.

**Step 2:** The **Semantic Distance** $SD$ between the topic and all other concepts in ontology is computed using the following formula,

$$SD = |d|_{C_i,C_j}$$ (4-1)

Here, $d$ is the number of edges or links between any two concepts, $C_i$ and $C_j$. The concepts in ontology are the terms that belong to a domain (or a particular subject).

**Step 3:** **Semantic Relevance** between two concepts of domain ontology is,

$$SR_{C_i,C_j} = \frac{1}{|d|_{C_i,C_j}+1}$$ (4-2)

Thus, Semantic Relevance is inversely proportional to the distance between any two concepts in ontology.

**Step 4:** **Dynamic Semantic Relevance**, $DSR_{P_i}$ of a web page, $P_i$, with respect to a topic ($C_i$) is calculated by summing up the product of the frequency of each term (from the expanded topic list) in the web page and its Semantic Relevance $SR_{C_i,C_j}$ (Eq. 4-2). This is formalized as following.

$$DSR_{P_i} = \sum_{j=1}^{n} \frac{f_{C_i}}{|d|_{C_i,C_j}+1}$$ (4-3)

$$DSR_{P_i} = \sum_{j=1}^{n} \left(f_{C_i} \times SR_{C_i,C_j}\right)$$ (4-4)

Here, $n$ is the total terms (concepts: $C_j$) in the expanded topic list and $f_{C_i}$ is the frequency of a concept $C_j$ divided by the total terms (excluding stop words) in a web page $P_i$. 
The complete procedure used by the proposed SFC is summarized in Algorithm 4-1. It explains various data structures used during the crawl along with the crawling procedure.

**Algorithm 4-1: Semantic Focused Crawler**

\[ \rho Q \rightarrow \text{priority queue containing URLs and their Dynamic Semantic Relevance,} \]

\[ DSR_{P_i} \]

\[ \xi Links \rightarrow \text{queue containing traversed URLs during the crawl to avoid cycles, thus it checks for duplicate traversals,} \]

\[ \epsilon T \rightarrow \text{expanded topic list consists of related terms (concepts) and their semantic distance to the topic from the ontology.} \]

Thus,

\[ T = \{t_0, t_1, t_2, ..., t_m\}, \text{ where } m > 0, t_0 \text{ is the topic for focused crawl.} \]

\[ \epsilon T = \{(c_0,0), (c_1,d_1), (c_2,d_2), ..., (c_n,d_n)\}, \text{ where } n > 0, \ c_i \rightarrow \text{semantically related to } c_j \]

and

\[ d_i \rightarrow \text{semantic distance.} \]

1. Initialize \( \rho Q \) with seed URLs
2. Repeat till \( !\rho Q.\text{empty()} \|| \text{fetch\_cnt} <= \text{Limit} \) {
   3. \( \text{web\_page.url} = \rho Q.\text{top.getUrl()} ; \) \hspace{1cm} // single url
   4. fetch and parse \( \text{web\_page.url} \);
   5. \( \text{web\_page.urls} = \text{extract urls (hyperlinks) from } \text{web\_page.url} ; \) \hspace{1cm} // list of urls
   6. for each \( \text{web\_page.urls} \) {
      7. \( \text{already\_exist} = \text{check } \text{web\_page.urls}[i] \text{ in } \xi Links ; \) \hspace{1cm} (duplicates)
      8. if \( !\text{already\_exist} \) {
         9. \( \text{enqueue } \text{web\_page.urls}[i] \text{ in } \xi Links ; \)
      10. \( \text{fetch and parse } \text{web\_page.urls}[i] ; \)
      11. \( \text{compute } DSR_{P_i} \text{ of } \text{web\_page.urls}[i] ; \)
12. enqueue \((web\_page\_.urls[i], DSR_{p_i})\) in \(Q\);

13. store \((web\_page\_.urls[i], DSR_{p_i})\) in local database;

14. } // end of if

15. } // end of for each

16. } // end of repeat

4.5 Discussion

This chapter discussed the need and design issues of the Focused Crawlers. It also proposed the design of crawling approaches for Semantic Focused Crawler and Social Semantic Focused Crawler.

The Social Semantic Focused Crawler called FCHC was proposed which used two crawling patterns BFP and DFP. Further, DFP was implemented at two different levels, level-1 and level-2. FCHC made use of bookmarked (tagged) web links on social web site and semantic knowledge to prioritize the sequence of web page traversal. The page relevance was computed based on the popularity of the web pages and tags that are assigned by the web user community usually by analyzing the web page.

Another crawling design was proposed for a multithreaded Semantic Focused Crawler (SFC). This crawler was used to fetch semantically relevant web pages from the Web on a given topic. The SFC used Dynamic Semantic Relevance (DSR) to prioritize the web pages to be crawled further. DSR was computed during the crawl for each web page, based on the expanded list of the topic and the semantic distances among various semantically linked concepts from the domain ontology. Domain ontology was constructed manually on a few learning subjects, to include most of the related concepts which were linked based on their semantic relations. The potentially relevant web pages found by the SFC were stored in a local database.

Although in comparison to the Semantic Focused Crawler, the Social Semantic Focused Crawler retrieved relevant results without incurring the overhead of parsing the content of each web page, yet in addition it required pursuing deep web
search on social portals which made the retrieval system dependent on the credibility of such sites. Also, only a small fraction of the Web which was bookmarked by the user community was accessed by the crawler. Probably, in the near future, a single point of access to various social network systems could make FCHC approach more beneficial as the search area and number of users would get assimilated.