6. Informal e-Mentoring using Multi Agent System

E-Mentoring is a process of knowledge dissemination by a more knowledgeable person (called Mentor) to a less knowledgeable person (called Mentee) through utilizing electronic and web technologies. In an informal e-Mentoring scenario the conversation between mentor and mentee takes place spontaneously, usually for a short duration, till the immediate goal of the mentee is achieved. Also, it does not require any prior investigations on the mentor-mentee relationship, being this mentoring informal. Informal e-Mentoring may use Virtual Mentors to provide knowledge to mentees.

This chapter presents a web based Multi Agent Framework that incorporates a Virtual Mentor who process a mentee’s information need, retrieves relevant information by communicating to various co-operative semantic information gathering agents, and deliver the relevant content to the mentee. The Framework makes use of crawling and ranking techniques which in turn makes use of the Concept Ontology for the mentoring purpose. A working prototype towards Virtual Mentoring integrated with many cooperative agents has been developed, based on the proposed framework.

6.1 Introduction

Mentoring, in general is defined as dissemination of knowledge in all kind of psychosocial and career related support by a more knowledgeable person to a less knowledgeable person (O’Neill, Weiler, & Sha, 2005), (Williams & Kim, 2011) regardless of age or status (Murphy & Ensher, 2006), and following the ethics and etiquettes guidelines (Daniel, 2006).

The more knowledgeable person who acts as a guide, coach, advisor, knowledge provider is called a mentor, whereas a less knowledgeable person who receives guidance to enhance his /her knowledge is called mentee or protégé. However, the term ‘mentee’ differs from the term ‘protégé’ by the way of mentoring provided to them. Mentee refers to a broad range of individuals who may be in the
role of ‘learner’ in mentoring relationships, irrespective of the mentor’s age or position whereas a protégé is identified as ‘under the wing’ of a mentor who is protected and nurtured over time (Daniel, 2006).

A new research area in the field of mentoring that is evolving in recent times is e-Mentoring (Haggard, Dougherty, Turban, & Wilbanks, 2011). E-Mentoring is considered as a technology oriented mentoring (Switzer, 2009) with an electronic dimension, which overcomes the challenges of face-to-face mentoring such as time and distance constraints, interpersonal incompatibility issues, relationship sensitivity etc.

Within an academic setting, where a mentee is in the role of a learner, mentoring can happen between advisor-student, from student to student in the form of knowledge transfer, instructor training or at peer faculty level (Mullen, 2007), (Williams & Kim, 2011), (Shrestha, May, Edirisingha, Linsey, & Burke, 2009).

The emergence of e-Mentoring owes to the arrival of internet and the supporting software. There exist few papers in literature on the topic of e-Mentoring since its emergence in late 1990s. Out of these, most of the papers have explored possibilities of applying computer technology to the existing research in the field of face-to-face mentoring. This resulted in the blended approach of Face-to-Face and Computer Mediated Communication (CMC) (Williams & Kim, 2011); (Ensher, Heun, & Blanchard, 2003), Telementoring or CMC only (DiRenzo, Linnehan, Shao, & Rosenberg, 2010); (Headlam-Wells, Gosland, & Craig, 2006). In this scenario the human mentors are allowed to communicate with mentees synchronously or asynchronously through computer aided devices along with simple programs to manage their communications. This type of e-Mentoring is more suitable for formal mentoring. The following sections present different types of e-Mentoring that may happen in different environments.

6.1.1 Computer Mediated Communication and Virtual Mentoring

A careful analysis of mentoring tasks suggests that among them many tasks can be accomplished through the use of some collaborative systems which may act as a Virtual Mentor. These Virtual Mentors (Zhang D., 2004); (Landowska, 2010) can
be deployed in a system to assist mentees in seeking information according to their needs, and answering basic queries intelligently and immediately, with 24x7 availability. The more peculiar problems such as those related to psychosocial mentoring may be transferred to the human mentors, for which they could then be able to spare more time and efforts to appropriately advice mentees.

### 6.1.2 Formal and Informal e-Mentoring

Mentoring can be formal or informal (Daniel, 2006). Informal mentoring is spontaneous and usually comes into effect to serve the need of a mentee on a specific task (Karkoulian, Halawi, & McCarthy, 2008). In contrast the formal mentoring is highly structured, implemented and monitored with proper planning for the mentor-mentee long term relationship compatibilities, establishment and measurable assessment of goals achievements etc. Both formal and informal mentoring has their benefits in the learning outcomes of a mentee (Singh, Bains, & Vinnicombe, 2002). However, in some cases such as, seeking answers to small and simple queries related to some information or knowledge needs, incompatible and strained relationship between mentor-mentee for a long term commitment or an actual time allotment disagreements are a few to be exemplify where the informal mentoring is beneficial to a mentee. Though informal mentoring may suffer from lack of an explicit framework within which it operates, still informal schemes have shown more success than formal schemes due to a number of factors related to mentoring other then listed above, namely making personal requests to mentees during mentoring, hesitance in disseminating knowledge and skills, lack of interaction time, gender biasness etc. (Singh, Bains, & Vinnicombe, 2002); (Karkoulian, Halawi, & McCarthy, 2008). Irrespective of the types of mentoring, the evidences from the past and ongoing research in the field of mentoring shows that people have been benefited from the mentoring programs (Allen & Eby, 2008); (Williams & Kim, 2011).

### 6.2 Related Work

The initial e-Mentoring program mentioned in literature (Single & Single, 2005) was Electronic Emissary Project (Harris J., 1994). It was aimed to mentor school children by connecting them to scientists, who mentored them in science
related projects. The next large project was Telementoring Project (Bennett, Tsikalas, Hupert, Meade, & Honey, 1998) that focused on gender equity in technology by pairing female high school students with female professionals in technology. It was an effort to encourage and support female high school students so as to inculcate their interest in the use of computers. The other two large scale projects were International Telementoring Program, founded by Neils (1997) and MentorNet founded by Muller (1997) are still serving the society by connecting mentors and mentees. The above mentioned pioneer projects and the followed up research in the same area majorly focused on the efficacy of building better mentoring relationships online and facilitate them with variety of synchronous and asynchronous user-friendly communication tools (Headlam-Wells, Gosland, & Craig, 2006); (DiRenzo, Linnehan, Shao, & Rosenberg, 2010). In addition to the mentor-mentees’ pairing, other mentoring tasks were explored like providing training, coaching and support to mentee, assessment (Akin & Hilbun, 2007), making available mentoring and learning resources (Headlam-Wells, Gosland, & Craig, 2006), effective use of communication tools (Williams & Kim, 2011), counseling and learning (Ensher, Heun, & Blanchard, 2003). All these e-Mentoring tasks involve human mentors, and in some cases even a teacher or coordinator who is represented as a mediator to the mentor and mentee.

Similar to the virtual mentor as proposed in the work, the literature in e-Learning witnesses the systems that make use of software agents (Gregg, 2007), (Agarwal, Deo, & Das, 2004) to perform various e-Mentoring tasks. For example, the use of Personal Agents for proactive and adaptive learning (Rosic, Stankov, & Glavinic, 2002), Pedagogical Agents to provide collaborative (Zhang, Soh, Liu, & Jiang, 2005) and individualized learning (Johnson, Rickel, & Lester, 2000), Content and Chat Agents for knowledge management tasks and user interaction (Bentivoglio, et al., 2010) resp., Emotional Agents to analyze facial expression of a learner and respond accordingly (Ammar & Neji, 2006) and Virtual Mentor Agents (Zhang D., 2004) are few of the contributions in the literature.

Mentees may have more than one mentor throughout their career. With multiple mentors, a mentee can benefit in different fields with diverse experiences and skills (Daniel, 2006). Today’s Web can be seen as a great source of information pool
where several experienced people write about their experiences and share their knowledge on different concepts and issues. A mentee, through such a system can be benefited by consuming these enriched resources. The main challenge here is to identify and understand the request of mentee and then retrieve relevant resources from a dynamic open system, the Web. The proposed framework is an effort to build an open system by incorporating above mentioned needs. This system would enable mentees to interact informally with virtual mentors for their informational needs.

6.3 Multi Agent Framework for Informal e-Mentoring

The proposed framework is an effort to automate the knowledge dissemination task of informal e-Mentoring. The framework makes use of the cooperative multi agents which in collaboration assimilate relevant Web content on the topics of users’ interest. This content is automatically augmented to the Concept Ontology by an agent to enable semantic retrieval of the content.

The overall purpose of the collaborative agents is to assist a Mentee by providing latest and useful information on a topic from the WWW and the Semantic Web (see chapter 2 for Web descriptions), especially when the Mentee is new to that

Figure 6-1: Framework for Informal e-Mentoring
topic. In such situations, the system does not only provide the related concepts on a topic to the Mentee, but also makes available the relevant web links and content from the Linked data (the published Semantic Web data).

Figure 6-1 illustrates the framework of the proposed system which demonstrates the background working on collecting information from the Web and building a knowledge base to finally disseminate knowledge to a Mentee. The sub-sections describe each agent illustrated in the Figure 6-1 and the tasks performed by them. The prominent communication scenarios among the agents have been also discussed.

6.3.1 Mentoring Task Force

The proposed framework as shown in (Figure 6-1) consists of several types of agents which are created by a few main agents such as PAM, VMA and so on, automatically in order to handle multiple activities associated with knowledge acquisition and mentees’ requests. Description of the responsibilities and accountabilities associated with each of the system’s main agent is given below.

**PAM- Personal Agent to Mentee:** Assists a mentee (learner) by passing search request to Virtual mentor (VMA) for relevant information and presenting it to mentee in appropriate presentable form. PAM being fundamentally a personal agent to a mentee, always remains active and assists a learner by recommending related concepts to the topic of his/her interest.

**VMA- Virtual Mentor Agent:** On request of PAM, it accumulates relevant information from other information gathering agents and returns back to PAM. It has the responsibility of creating a personal PAM for each mentee on the system. It is also liable to create and assign a personal mentor agent called MentorForMentee. This agent is alive until the immediate information need of a mentee gets achieved. The agent MentorForMentee is thus a personal Virtual Mentor to a mentee. This provision is primarily required to keep the VMA at the listening mode so as the request of each mentee in the system can be responded in time. The MentorForMentee agent principally communicates with the OntoAgent and DbpediaAgent to gather relevant content on the mentee’s topic of interest.
**DbAgent**: is an agent responsible for managing the downloaded web resources and its associated data. It makes available the data on request of other agents. DbAgent remains alive throughout and chiefly communicates with the OntoAgent and the Web Crawler.

**OntoAgent**: is an ontology agent. It manages the Concept Ontology and updates the knowledge base as and when required. Besides this, it provides services such as Term Expansion on request of other agents. OntoAgent also select concepts from the Concept Ontology and send to DbAgents which are then crawled, and resources are collected, through the Crawler Agents.

**DbpediaAgent**: is an agent that can access the exact information related to a given concept from the dbpedia, the linked data. The working of the agent shows that how software agents could read the linked data to serve semantic web enabled applications. This agent can be extended to read other semantic information on the Semantic Web.

**Crawler Agents**: Semantic and FCHC crawler agents crawl the Web for relevant resources. Semantic Focused Crawler (SFC) crawls the WWW whereas Focused Crawler based on Human Cognition (FCHC) crawls a SBS to collect web resources on the topics/concepts (crawling has been described in chapter 4 in detail) provided by the DbAgent. The crawlers, after completing their task contact to DbAgent with the collected data.

The communication among agents is handled asynchronously. In order to minimize the response time and to handle multiple requests from various sources (users or agents), the main agents of the system create assistant agents and designate subtasks to them. In the meantime the main agents process other request in the queue.

### 6.3.2 VMA serving Mentee’s request

Figure 6-2 shows the complete sequence of how a mentee’s request is processed by the VMA and then a response is send back. Following are the steps followed to achieve this task.
Step 1: A mentee enters a topic term with its domain and level of information (Basic, Average or Advanced) to get relevant information.

Step 2: PAM gets activated with this entry and passes this information to the VMA.

Step 3: VMA creates an instance of its assistant (MentorForMentee) to serve individual PAM.

Step 4: MenterForMentee agent communicates with OntoAgent, DbpediaAgent to get information.

Step 5: This information is communicated back to the requesting PAM through VMA.

Step 6: PAM creates a presentable web page to impart extracted knowledge to its Mentee.

6.3.3 DbAgent processing crawling task

Figure 6-3 shows the sequence of agents’ communication involved in the process of ontology augmentation. As description in chapter 3, ontology augmentation
is the adding up of the potentially relevant web resources to the ontology and linking up these web resources to the ontology concepts according to their relevance. Ontology augmentation thus helps in generating the knowledge base automatically.

Following steps describe the process in detail.

**Step 1:** DbAgent initiates its process on request generated by OntoAgent with the provided information (Topic and expanded topic structure which contains all required semantics for resource retrieval).

**Step 2:** DbAgent starts the instances of focused crawler agents, called SemanticAgent and FchcAgent to search each assigned topic while keeping the record of all sql tables.

**Step 3:** Focused crawlers on finishing the assigned task intimate to DbAgent.

**Step 4:** DbAgent computes the relevance of each web resource.

**Step 5:** DbAgent communicates back to OntoAgent with the information regarding resource tables.
6.3.4 OntoAgent communication with other agents

OntoAgent primarily communicates with VMA or DbAgent. OntoAgent interacts with VMA (Figure 6-2) to provide required the semantic information whereas it communicates with DbAgent (Figure 6-3) to update the Concept Ontology. The former sequence has been described in section 6.3.2. Following steps details the latter sequence of the communication.

Step 1: For each existing domain in the knowledge base, OntoAgent creates an instance of a DomainAgent with the name of its domain.

Step 2: Each DomainAgent sends the search request for each of its concept one by one, to DbAgent through OntoAgent.

Step 3: After getting its request served by DbAgent, it updates its domain ontology by augmenting relevant resources to the topic (concept).

6.3.5 DbpediaAgent to access the Linked Data

DbpediaAgent as illustrated in Figure 6-2, accesses the Linked Data to retrieve information on a topic requested by the agent MentorForMentee, which is created by VMA. To access the Linked Data, DbpediaAgent creates an online connection through the SPARQL EndPoint service\(^\text{34}\). The DbpediaAgent performs the following steps to retrieve content from the Dbpedia of the Linked Data.

Step 1: DbpediaAgent receives a topic from the request generated by the virtual mentor.

Step 2: DbpediaAgent creates a query by itself using jena query factory API.

Step 3: It makes an online connection through SPARQL EndPoint service to execute the query.

Step 4: The returned content from the Linked Data (Dbpedia) is then sent to the corresponding Virtual Mentor (MentorForMentee) for further processing.

There could exist similar agents working on different semantic data contained in the Linked Data Cloud. In that case the Virtual Mentor should be able to retrieve the required information with ease for mentoring purpose.

---

\(^{34}\) http://dbpedia.org/sparql
6.4 Experimental Study and Evaluation

The experimental study has been conducted in three phases. Each phase consists of the implementation of the components and their evaluation/verification depending on the evaluation needs. The first phase focuses on the construction and verification of the Concept Ontology based on the design described in chapter 3. In the second phase various focused crawlers have been implemented and evaluated for their crawling efficiency. The designs of these focused crawlers were presented in Chapter 4. The third phase finally evaluates web resources’ ranking (ranking concepts presented in chapter 5) recommended by the system for the user satisfaction. The experimental study, evaluation metric or tool and, the results obtained from the study are detailed below.

6.4.1 The Concept Ontology

![Screen shot of the Protégé editor displaying the Concept Ontology](image)

Figure 6-4: Screen shot of the Protégé editor displaying the Concept Ontology
The Domain Ontology has been constructed on three domains: database, Computer Network and Java programming Language using Protégé editor. The Figure 6-4 shows a screen shot of the protégé editor, which displays a collation of various components of the Concept Ontology. The Concept Ontology presently consists of 14 Classes, 21 Object properties, 4 Data properties and more than 50 individuals. There are 8 sub class axioms and 6 sub object property axioms. The ontology also has inverse, functional, inverse functional and transitive axioms on the object property other than the domain and range axioms.

The Concept Ontology has also been validated on OWL Ontology Validator WonderWeb\(^{35}\), which showed a successful validation for OWL DL and OWL Full. This ontology has been used for the next two phases of the experimental study and evaluation.

### 6.4.2 The Focused Crawlers

Design of the proposed crawlers, FCHC (Focused Crawling based on Human Cognition), a social semanticFocused Crawler and DSRbasedSFC (Dynamic Semantic Relevance based SFC), a Semantic Focused Crawler (SFC) have been explained in the chapter 4. These crawlers have been executed on various query topics and compared with a basic focused crawler named as classic focused crawler. The DSRbasedSFC crawls on content semantic similarity of the web pages and the FCHC crawls on the semantic relevance of bookmarked tags.

#### 6.4.2.1 Setup

The coding on the proposed approach has been mostly written using java programming language, and Protégé OWL APIs and Jena to access ontologies. The knowledge base consisting of ontologies has been developed using Protégé (explained in section 2.4.3). The SBS crawler was coded using java. MySql has been used to store crawled relevant resources. The experiment was conducted using three topics from different domains. The crawlers were executed on Intel core 2 Duo processor, 2.4 GHz, 2 GB RAM, 32-bit OS, for over 30 times with two distinct sets of seed URLs on each of the three topics. The topics were expanded using domain ontologies.

---

\(^{35}\) http://www.mygrid.org.uk/OWL/Validator
Semantic distance of each expanded term from the topic term was also computed.

6.4.2.2 Implementation

The experiment was conducted on sets of queries from three domains (listed in Table 6-1). Depth of the resource content required by a user was assumed as of ‘Basic level’, which means all super-concepts and single level sub-concepts of the topic would be added to the expanded list from the domain ontology.

<table>
<thead>
<tr>
<th>Domains</th>
<th>Query terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>java programming language</td>
<td>java, datatype, variable, loop, applets, event handling, recursion</td>
</tr>
<tr>
<td>database</td>
<td>dml, sql, insert into, set operators, group by</td>
</tr>
<tr>
<td>computer network</td>
<td>transmission media, integrated circuits, coaxial cable, topologies, TCP/IP, protocols, modem, subnet</td>
</tr>
</tbody>
</table>

This approach can be implemented for any topic, provided the domain ontology of the query term exists. To conduct the experiment three domain ontologies were constructed, and therefore the queries were purposely chosen from these domains.

**FCHC:**

The focused crawler FCHC was implemented to download bookmarks from the SBS, delicious.com. The top 10 search results from the search engine (Google) on the query terms were fed as seeds to initiate crawl in SBS. Finally, the proposed approach SSR was executed on the potentially relevant bookmarks downloaded by the crawler. It computed the social semantic relevance between each of the downloaded resource and the search query using similarity measure cosine 0 as explained in the previous section. The resources were then arranged in the descending order according to their computed relevance (Cos 0) in order to rank them.

The experiments related to FCHC focused crawlers were performed in two parts. In the first part the comparison was studied among three variations of the proposed FCHC focused crawlers based on their different crawling search patterns.
The social information for these crawlers was computed from the SBS combined with the semantic knowledge from the domain based Concept Ontology. As already mentioned in section 4.4.1 these focused crawlers were made to crawl the SBS. Each of these FCHC crawlers have used one of the three patterns viz. Breadth First Pattern (BFP), Depth First Pattern (DFP) up to level 1 and DFP up to level 2. Although, there exist a number of SBS, viz. digg\textsuperscript{36}, StumbleUpon\textsuperscript{37}, reddit\textsuperscript{38}, delicious\textsuperscript{39}, DZone\textsuperscript{40} which can be used for crawling by designing appropriate parsers, for the experiment purpose we have used the largely accessed bookmarking site, delicious.com. In a more demanding state, many crawlers can be built to crawl multiple sites simultaneously to collate the social information which effectively increases the exploration area on the Web.

The other part of the experimental study was used to compare FCHC results with the results obtained from Classic Focused Crawler and Semantic Focused Crawler. The three focused crawlers including the Social Semantic Focused Crawler (DFP-L2) differ in their selection for the resource relevance criteria and the crawling web area.

The Classic Focused Crawler is a baseline crawler that uses hyperlinks on the web page to crawl. It initiates the crawling with a set of URLs. Each web page is parsed by the focused crawler. During the page parsing, only relevant links (URLs) are put in the queue. Just to recall here, the links are considered relevant, in particular to the classic focused crawler, if the parsed anchor text contains the search topic. The priority criterion for the Classic Focused Crawler is FCFS (First Come First Serve). The topic relevance of each crawled document is computed later using the same Semantic Relevance. The Semantic Relevance is computed by summing up the semantic distance of all semantically related words present on the web page. If the page is semantically relevant, it is put in the queue for further crawl. The priority of the page selection for the next crawl is decided depending on the computed Semantic

\textsuperscript{36} digg.com/
\textsuperscript{37} www.stumbleupon.com/
\textsuperscript{38} www.reddit.com/
\textsuperscript{39} http://delicious.com/
\textsuperscript{40} www.dzone.com/
Relevance.

The classic FC and Semantic FC search resources on the WWW, whereas the Social Semantic Focused Crawler: FCHC-DFP-L2 and its other variants crawl the SBS to collect potentially relevant resources.

**DSRbasedSFC:**

The semantic focused crawler, DSRbasedSFC was implemented with five threads to gather resources from the WWW. The design of the crawler has been presented in the section 4.4.2. It uses domain ontology that contains the topic to be crawled. The topic is expanded with semantically related terms using the ontology. Based on this expanded topic list and the semantic distances among various related concepts, the multithreaded SFC fetches and parses web pages to compute the dynamic semantic relevance.

A priority queue is maintained containing web pages and their dynamic semantic relevance to prioritize further crawls. Each thread picks the highest priority (top link in priority queue) URL to parse all links on the web page and put them again in the queue with their dynamic semantic relevance. They also store relevant web pages in the local database. A separate queue consisting of all traversed URLs to keep a check on duplicate or mirror links is also maintained in order to avoid crawling cycles.

An open source multithreaded java crawler called crawler4j\(^{41}\) was used as a baseline focused crawler to compare the results of the proposed DSRbasedSFC crawler. It was customized to incorporate topic and seed URLs by overriding the methods shouldVisit(WebURL url) and visit(Page page). The results of both crawlers were compared by taking same set of seed URLs, topic and number of threads. The difference between the designs of two crawlers lies in the priority computation based on semantic relevance. Crawler4j is considered as the baseline focused crawler without any priority check for the experimental purpose, whereas the proposed crawler SFC uses dynamic semantic relevance to prioritize the visiting sequence of

\(^{41}\) http://code.google.com/p/crawler4j/
web pages. All relevant URLs are stored in MySql database during the crawl.

6.4.2.3 Evaluation Metric

The web resources collected through various crawlers were evaluated using the same evaluation metric to measure semantic similarity between the resources and the queries. A separate program computes semantic relevance of the fetched web pages, based on this metric. This program fetches the URL tuple stored during the crawl from the database and compute their semantic similarity. The method to compute semantic similarity is described in the following sub-section.

A user adjustable threshold is used to filter relevant web pages. The threshold may include one of the values from the range (”>0”, ”≥0.2”, ”≥0.4”, ”≥0.6”). The value of the threshold is adjusted manually so that nearly 50% of top relevant web pages can be considered for experimental evaluation.

This helps to evaluate focused crawlers more rigorously, as the high threshold implies high value of relevance for a web page. The crawled results of both crawlers have been evaluated using semantic similarity measure based on Vector Space Model. The metrics used in evaluation are briefly explained below.

*Semantic similarity:*

In order to evaluate the two crawlers for the experiment, semantic similarity between crawled web pages by the crawlers, and the expanded topic list have been considered. The semantic similarity measure is based on the Vector Space Model, where the term weights in the model are computed from the Concept Ontology. is the weights have been calculated from the semantic distance among concepts in the ontology. These are weighted according to the frequency of each semantically relevant term in a web page, which thereafter are used as vectors to compute cosine similarity. The vector lengths\(^{42}\) of a topic vector and a web page vector are computed using following methods:

The topic vector length \(|T|\) is computed using the semantic distance between the query term and it’s each expanded term as the following function:

\(^{42}\) A vector consists of two components, a direction and magnitude (length). In information retrieval direction has no meaning therefore, only length is used for computational purpose.
|T| = \sqrt{\sum_{i} Wt_{t_{0}-t_{i}}^2} \quad (6-1)

Here, $Wt_{t_{0}-t_{i}}$ is the weight assigned to each $t_{i} \in T$, \forall m terms. It is formulated as following:

$$Wt_{t_{0}-t_{i}} = \frac{SR_{t_{0}-t_{i}}}{\sum_{i=0}^{m} SR_{t_{0}-t_{i}}} \quad (6-2)$$

$SR$ is the Semantic Relevance computed using the Domain Ontology that contains a topic.

Similarly, a web page vector length is computed for each web page, using the following function:

$$|P_{i}| = \sqrt{\sum_{k} Wt_{p_{k}-p_{i}}^2} \quad (6-3)$$

$Wt_{p_{k}-p_{i}}$ is the weight assigned to each term $p_{k} \in T$ that exist in the web page $P_{i}$ so that,

$$Wt_{p_{k}-p_{i}} = \frac{SR_{t_{0}-p_{k}} \cdot f_{p_{k}}}{\sum_{k=1}^{m} (SR_{t_{0}-p_{k}} \cdot f_{p_{k}})} \quad (6-4)$$

The normalized cosine similarity between $t_{0}$ and a web page $P_{i}$ is defined as:

$$\cos \theta_{t_{0}-p_{i}} = \frac{T \cdot P_{i}}{|T| \cdot |P_{i}|} \quad (6-5)$$

where, $T \cdot P_{i}$ is the dot product computed as,

$$T \cdot P_{i} = \sum_{j=p_{k}} (Wt_{t_{0}-t_{j}} \cdot Wt_{p_{k}P_{i}}) \quad (6-6)$$

Therefore, using Eqs. (6-1), (6-3), (6-5) and (6-6), we get,

$$\cos \theta_{t_{0}-p_{i}} = \frac{\sum_{j=p_{k}} (Wt_{t_{0}-t_{j}} \cdot Wt_{p_{k}P_{i}})}{\sqrt{\sum_{i} Wt_{t_{0}-t_{i}}^2} \cdot \sqrt{\sum_{k} Wt_{p_{k}P_{i}}^2}} \quad (6-7)$$

The value of similarity measure ($\cos \theta$) is maximum for the most relevant web page and zero for irrelevant web pages. Each web page is thus evaluated for its similarity with the topic.
Harvest ratio:

Harvest ratio \( (hr) \) (Chakrabarti, Berg, & Dom, 1999), (Batsakis, Petrakis, & Milios, 2009), (Menczer, Pant, Ruiz, & Srinivasan, 2001), (Zheng, Kang, & Kim, 2008) is defined as the fraction of web pages crawled that satisfy the crawling target (relevant pages) \#r among the total crawled pages \#p. Thus \( hr = \frac{\#r}{\#p}, hr \in [0...1] \). The relevance of the crawled web pages is determined by computing semantic similarity metric described above.

6.4.2.4 Results

**FCHC based focused crawler results:** Figure 6-5 shows the total number of web resources crawled by the crawlers under experimental study. There are three bars represented by each focused crawler in the graph. The first bar shows total resources crawled whereas second and third shows the relevant resources having relevance threshold greater than zero and, greater than or equal to 0.2 respectively. The relevance threshold here represents the relevance value of a resource, above which the resource may be considered relevant to a given search topic. It is apparent from the Figure 6-5 that a more number of web resources are found while going deep in the social portal. However the ratio of potentially relevant web resources to the total crawled web resources reduces. At the same time, the BFP and DFP-L1 gives a high ratio, but overall less number of web resources. The SFC and Classic FC gathered
web resources from the WWW instead of a social portal. The SFC showed better performance over Classic FC, BFP and DFP-L1. But overall DFP-L2 performed best over other focused crawlers. A more detailed analysis on this experiment can be obtained by studying following graphs.

**Comparison among variants of FCHC crawlers:** The FCHC crawler (Figure 6-5) with BFP approach collected 1108 resources in 1hr 49 min; DFP-L1 collected 774 resources in 1 hr and DFP-L2 collected 22,552 resources in 20 hrs 9 min. Though DFP-L2 took longest time to collect resources, it retrieved maximum (13584) number of relevant resources when compared to BFP (688) and DFP-L1 (513) with similar input data. Figure 6-6, 6-7 and 6-8 show the harvest ratio of the resources crawled by FCHC variants viz. FCHC-BFP, FCHC-DFP-L1 and FCHC-DFP-L2 crawlers respectively.

![Figure 6-6: Harvest Ratio of FCHC-BFP crawler](image)

The structure of the SBS is such that the crawler is able to achieve better results when the resources are traversed in Depth first pattern and it is best when the search is deeper. The results (Table 6-2) show a slight better harvest rate of BFP and DFP-L1 than the DFP-L2 over the complete data set. This show that the difference between the relevant resources with relevance threshold >= 0.2 and the total crawled resources are lesser in the earlier two cases, but those do not cover a large specified area on the Web.
Comparison among FCHC and other crawlers: Figures 6-8, 6-9 and 6-10 show the harvest ratio of the resources crawled by FCHC-DFP-L2, Semantic FC and Classic FC crawlers respectively. All graphs show the harvest rate starting from 1 to downwards and then again rising upwards to finally reaching to a stable rate. The initial high values are due to the seed URLs, out of which most are relevant to the search topic.

Classic Focused Crawler shows the worst downward graph for not able to find or analyze relevant resources in the Web. This is definitely due to the missing social and semantic information regarding the search topic. Semantic Focused Crawler gives better performance than the Classic Focused Crawler; however, FCHC-DFP-L2
shows the best harvest ratio reaching above 0.8. This clearly shows the significance of using semantic information by the Semantic Focused Crawler and above it social and semantic knowledge together in FCHC-DFP-L2.

Harvest rate of FCHC and other focused crawlers: Figure 6-11 shows the harvest rate of the crawled web resources by all five focused crawlers. Harvest ratio (hr) is the fraction of web pages crawled that satisfy the crawling target (relevant
Informal e-Mentoring using Multi Agent System

Pages) \( #r \) among the crawled pages \( #p \). Thus, \( hr = \frac{#r}{#p} \), \( hr \in [0...1] \) and harvest rate is the harvest ratio per unit of time. FCHC-DFP-L2 shows the best stable harvest rate among all the crawlers.

![Figure 6-9: Harvest rate of crawlers under study](image)

Table 6-2 shows the harvest ratio of the five crawlers with a threshold above zero and above 0.2. A high ratio of FCHC-DFP-L1 shows that over other crawlers, it gathered lesser irrelevant resources out of the total crawled web resources.

<table>
<thead>
<tr>
<th>Harvest Ratio with different threshold</th>
<th>Classic FC</th>
<th>Semantic FC</th>
<th>FCHC-BFP</th>
<th>FCHC-DFP-L1</th>
<th>FCHC-DFP-L2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( hr (RT &gt; 0) )</td>
<td>4.13%</td>
<td>85.27%</td>
<td>97.11%</td>
<td>99.35%</td>
<td>98.55%</td>
</tr>
<tr>
<td>( hr (RT &gt;= 0.2) )</td>
<td>0.42%</td>
<td>36.04%</td>
<td>62.09%</td>
<td>66.27%</td>
<td>58.79%</td>
</tr>
</tbody>
</table>

**DSR-based SFC focused crawler results:** The SFC crawler is evaluated using two distinct sets of seed URLs on each of the three topics (‘dml’, ‘transmission media’ and ‘integrated circuits’). The results were compared with an open source multithreaded focused crawler, Crawler4j using same data set and evaluation metric for both crawlers. Crawler SFC showed better harvest ratio for all the topics. Another evaluation indicator, robustness of a crawler to acquire relevant web pages related to a topic was also measured for SFC. The URL overlap ratio over the crawled resources was measured and was found to be as high as 0.8 for most of the topics, which shows
that the proposed crawler produces relevant results irrespective of different sets of the seed URLs.

Figure 6-12 shows the number of resources crawled by the DSRbasedSFC or simply SFC and the customized Crawler4j, called Classic FC. The input dataset for both crawlers is same. The comparison is made for unique number of crawled resources, and relevant resources with the threshold (semantic relevance value) greater than 0, and greater than 0.2. It shows that the Classic FC though crawled more number of unique web resources, more number of relevant web resources are however crawled by the SFC. The analysis infers the efficiency of the SFC which crawled 88% of relevant resources while Classic FC crawled 31% only.

![Figure 6-10](image)

**Figure 6-10: Number of resources crawled by SFC and Classic FC crawlers on same topic and on same set of the seed URLs**

**Comparison between (DSRbased)SFC and Classic FC:** Figure 6-13 shows harvest ratio of SFC and baseline classic focused crawler for the topic ‘dml’ at the threshold ‘>=0.4’. Initially starting from 1, the ratio first decreases and then picks up. The Initial high ratio is seen due to the seed URLs which are all relevant web pages. However all the links on those pages are not relevant, as a result of which the ratio comes down for a short period. Later the ratio picks up due to the better web page crawling. Figure 6-14 shows the harvest ratio of SFC and baseline classic focused
crawler for the topic ‘transmission media’ at the threshold ‘>=0.2’. It shows contrasting results between SFC and baseline crawler, with the SFC harvest ratio as good as 0.8. A harvest ratio comparison of SFC and baseline classic focused crawler for the topic ‘integrated circuit’ at the threshold ‘>=0.2’ is shown in Figure 6-15.

![Harvest Ratio for the topic 'dml'](image)

**Figure 6-11: Harvest ratio for the topic 'dml'

![Harvest Ratio for the topic 'integrated circuit'](image)

**Figure 6-12: Harvest ratio for the topic 'integrated circuit'
Robustness of DSRbasedSFC: Figures 6-16, 6-17 and 6-18 show the crawl robustness for topics ‘dml’, ‘transmission media’ and ‘integrated circuits’ respectively. Starting with zero overlap, it steadily rises up going beyond 0.8 URL overlap ratio which shows a remarkable robustness of the crawler for the ‘dml’ topic. The SFC crawlers are initiated with two distinct set of seed URLs (the reason for initial zero overlap) but after a short span of time both of the crawls show a number of common URLs as can be seen in robustness graphs. However, for the topic ‘integrated circuit’ the ratio declines after traversing approx 1500 web pages. The reason could be the addition of few web sites from the seed URLs list which may consist of a number of relevant links on them, would have been crawled by one of the crawlers.

Figure 6-13: Harvest ratio for the topic 'transmission media'

Figure 6-14: Crawl robustness for the topic 'dml'
6.4.3 Ranking of Web Resources

Any IR technique is evaluated by measuring its effectiveness (Sanderson & Braschler, 2009) i.e., by determining the relevance of retrieved items, relative to a user’s information need. The web resources obtained after crawling the social bookmarking site using the FCHC approach were evaluated by the experts (teachers or research scholars) of the domains selected for the experiments. Besides, an
exhaustive evaluation was carried out with the help of 31 non probabilistic purposive respondents, on randomly chosen three queries from the ‘database’ domain. The reason for choosing database domain was that the group of respondents had some basic knowledge on database subject. Ontology on database domain was built following the procedure as described in section 6.4.1, and the query topic ‘dml’ was expanded using the domain ontology (topic expansion has been described in section 3.5 and in chapter 5 while computing ranks of the web resources). For rest of the queries, experts from the respective domains were approached to evaluate the system generated ranks, as carrying evaluation process with such large groups is expensive in terms of time and resources. Nevertheless, the similarities in the evaluated results for select query terms by both types of the judges justify the evaluations done on other queries by experts.

More specifically, here we evaluate the results to check/measure:

- Whether the approach has able to collect relevant web resources from SBS and whether it shows any benefit of using collaborative tagging?
- Whether the rank order generated by SSR on crawled web resources correlates with the user’s relevance judgment?
- Whether the learners are satisfied with the ranking order generated by the proposed approach?

The experimental setup for creating test collection, and to evaluate the effectiveness of the proposed system is described in the following sub-sections.

6.4.3.1 Setup for Creating Test Collection

For creating the test collection, a questionnaire was designed to judge 15 web resources for their graded relevance to the query term. These web resources consist of top 15 resources ranked by SSR. The sample population was taken purposively consisting of 31 undergraduate students who were studying ‘Databases’ in their course and were beginners for the concept ‘dml’, which was also the query topic in the experimental study. The work in this paper is intended for the e-Learning domain, therefore to an extent the user’s intension while searching is assumed as the need for some basic understanding about the topic. This assumption led us to precisely define
the relevance using a concrete structure of domain knowledge. The assumption also supports the judges to be represented as the real searchers to an extent. After face to face introduction with respondents and training them regarding the search requirements and procedure to fill questionnaire, the communication was done through email. Each respondent was instructed to open web resources one by one which were represented by their URLs. They were suggested to browse and analyze each resource and then select one option from a four point scale, where each point represents most relevant, fairly relevant, marginally relevant or irrelevant respectively, similar to the criterion considered by Jarvelin & Kekalainen (Jarvelin & Kekalainen, 2002).

6.4.3.2 Relevance Judgments

Quality of the relevance judgment by humans variably differs by their capability. If they are given a set of web resources to rank them in some order, they would be able to do that correctly, possibly up to 5 or 6 ranks, after that the quality of their judgment deteriorate due to a number of reasons. They might not remember the contents of early resources while comparing them with later resources. In general, one is able to remember the contents up to a few numbers of initial resources only to put them in some order; beyond that, the chances of ranking them with wrong decisions increases. Alternatively, the judges can be asked to assign individual web resources a degree of relevance, e.g., 3, 2, 1, or 0 for highly relevant, fairly relevant, marginally relevant or irrelevant respectively. This type of a 4 point scale assessment allows a respondent to judge web resource relevance independent of other web resources in the list and allows the assessment of more number of web resources; without the need to relatively assess all resources. However, the assessment of results, in the presented experiment is conducted with both of the methods. An expert’s relevance feedback consisting of ranked web resources to the query relevance has been used in the first method; whereas the later assessment method uses the relevance judgment of 31 purposive respondents.

Moreover, the web resources are judged on the stated topical relevance, in the presented experiment. The term ‘stated relevance’ was originally given by Cleverdon & Keen (as stated by Ellis (1996)). It indicates that the resource relevance is judged
by a respondent other than the actual searcher while considering the given search conditions.

### 6.4.3.3 Evaluation Metric DCG(n): Discounted Cumulative Gain

DCG and nDCG (Jarvelin & Kekalainen, 2002) are useful metrics to measure the effectiveness of IR techniques as they combine the degree of web resources relevance and their ranks in a coherent way. It computes cumulated gain (Kekalainen & Jarvelin, 2002) as a single measure at any number of crawled web resources, irrespective of the recall base size. It is an improved measure over Cumulated Gain, motivated by the facts that:

- highly relevant documents are more valuable than marginally relevant documents, and
- greater the ranked position of the relevant document, the less valuable it is for the user, because less likely it is that the user will ever examine the document.

Accordingly, the discounted function progressively reduces the document score by the log of its rank, as its rank increases. Consecutively, for a greater rank, a smaller share of the document score is added to the cumulated gain. The cumulated gain vector with discount DCG is defined recursively as the vector DCG, where

\[
DCG[i] = \begin{cases} 
CG[i], & \text{if } i < b \\
DCG[i - 1] + G[i]/\log_b i, & \text{if } i \geq b 
\end{cases}
\]

Base of the logarithm is selected to make the discounts sharper or smoother. For example, \(\log_2 2 = 1\) and \(\log_2 1024 = 10\), which means that the document at the position 1024 would get one tenth of its gain value; whereas with base 10, i.e., \(\log_{10} 10 = 1\) and \(\log_{10} 1000 = 3\) would get one third of its gain value while documents positioned from 1 to 9 would get their usual cumulative gain.

**Gain vector, G:** Gain vector is the ordered list of relevance score called gain value for each document assigned by a judge. The gain values are arranged in order to the ranks obtained by the IR ranking technique. In this evaluation, relevance scores ranges from 0 to 3, where 0 represents irrelevant or no value and 3 represents highly
relevant or high value.

**Cumulated Gain, CG:** Cumulated Gain vector is defined recursively as the vector \( CG \) where:

\[
CG[i] = \begin{cases} 
G[i], & \text{if } i = 1 \\
CG[i - 1] + G[i], & \text{otherwise}
\end{cases}
\]

Thus, the cumulative gain at ranked position \( i \) is computed by summing from position 1 to \( i \) when \( i \) ranges from 1 to \( n \) (where \( n \) is the last rank in the list of relevant documents).

**Ideal DCG(n):** This vector represents theoretically best possible vector. It is constructed using the best possible score vector \( BV \) for \( k \), \( l \), and \( m \) relevant documents at the relevance levels 1, 2, and 3. The method of the construction is as follows.

\[
BV[i] = \begin{cases} 
3, & \text{if } i \leq m, \\
2, & \text{if } m < i \leq m + l, \\
1, & \text{if } m + l < i \leq m + l + k, \\
0, & \text{otherwise}.
\end{cases}
\]

An ideal gain vector obtained above is a sorted descending order of an actual gain vector. Ideal DCG vector is then computed in the same way as DCG(n) using the ideal gain vector.

**Normalized DCG(n):** The DCG vectors are normalized by dividing them by the corresponding ideal DCG vectors, component by component. Therefore for any vector position, the normalized value 1 represents ideal performance, and lies in the range (0, 1). Thus, formally we get the normalized DCG vector as,

\[
nDCG(n)[i] = \frac{DCG(i)}{IDCG(i)}
\]

The normalized DCG, nDCG is used to compare the results up to a given rank generated by different IR techniques, over a set of queries and thus gives a measure to analyze their performance differences.
6.4.3.4 Results

This section presents a detailed analysis and evaluation of ranked results, their comparisons with the resources retrieved by a search engine, and rankings given by human judges. The output of SSR on various queries showed overall a satisfactory performance. However, very few relevant documents were found for advanced concepts from almost all domains. On the other hand, a significantly large number of relevant documents were retrieved when queried on broader concepts.

Out of three domains, java_programming_language and database showed comparatively better retrieval of documents as compared to the Computer_Network domain. This observation infers that the concepts for which the community participation is active at a large scale, a larger set of very relevant documents were retrieved on those concepts. Nevertheless, the documents on even advanced concepts, though small in number, have been found very relevant to the search topics.

The SSR generated ranked results on the queries listed in Table 6-1 were evaluated by human judges. Results of ‘DML’, ‘SQL’ and ‘Group_by’ queries were evaluated by purposely selected respondents who were beginners in the database subject. Other queries were evaluated by 3 to 8 experts in each domain for each query. A detailed analysis is however presented for ‘dml’ query with few results on ‘sql’ and ‘group_by’ queries. Almost similar results were observed for queries from other domains.

Web Resources gathered from SBS: In reference to the above sited question for the evaluation of the presented approach, that whether the approach has crawled additional relevant web resources from SBS and whether it shows any benefit of using collaborative tagging, the answer is yes. Table 6-3 shows the comparison among ranks computed by SSR, search engine and the expert (from the field of the topic chosen for the experiment). Although in the presented experiment, Google is chosen as the search engine, however it can be replaced by any other search engine also. Following observations have been inferred and illustrated in Figure 6-19:
The result generated by SSR consists of 18 relevant resources out of 25 retrieved web resources (10 from search engine’s top result and rest from SBS),

4 resources being not bookmarked in SBS could not be evaluated by SSR; although among these 4 resources, 3 were not relevant from the expert’s view and one got the rank 16 from the expert, which shows comparatively very little relevance to the query topic,

Resources marked relevant by SSR approach were also found relevant by the expert and same was the case with irrelevant resources, except those 4 resources

<table>
<thead>
<tr>
<th>Web Resources (URLs)</th>
<th>Social Semantic Relevance</th>
<th>SSR ranks</th>
<th>Google order</th>
<th>Expert ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://en.wikipedia.org/wiki/Data_Manipulation_Language">http://en.wikipedia.org/wiki/Data_Manipulation_Language</a></td>
<td>0.805085</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><a href="http://www.tomjewett.com/dbdesign.php?page=ddl/dml.html">http://www.tomjewett.com/dbdesign.php?page=ddl/dml.html</a></td>
<td>0.801896</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td><a href="http://www.orafaq.com/faq/what_are_the_difference_between_dll_dml_and_dcl_commands">http://www.orafaq.com/faq/what_are_the_difference_between_dll_dml_and_dcl_commands</a></td>
<td>0.769261</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><a href="http://www.cs.sfu.ca/CC/354/zaiane/material/notes/Chapter1/node1.html">http://www.cs.sfu.ca/CC/354/zaiane/material/notes/Chapter1/node1.html</a></td>
<td>0.698888</td>
<td>4</td>
<td>*</td>
<td>3</td>
</tr>
<tr>
<td><a href="http://www2.aoao.gov.au/2dFGRS/Public/Release/Database/sql_intro.pdf">http://www2.aoao.gov.au/2dFGRS/Public/Release/Database/sql_intro.pdf</a></td>
<td>0.698888</td>
<td>5</td>
<td>*</td>
<td>4</td>
</tr>
<tr>
<td>databases.about.com/od/sql/a/sqlfundamentals_3.htm</td>
<td>0.646596</td>
<td>6</td>
<td>*</td>
<td>6</td>
</tr>
<tr>
<td>dev.mysql.com/doc/refman/5.0/en/insert.html</td>
<td>0.432221</td>
<td>7</td>
<td>*</td>
<td>8</td>
</tr>
<tr>
<td>dev.mysql.com/doc/refman/5.0/en/create-table.html</td>
<td>0.415603</td>
<td>8</td>
<td>*</td>
<td>14</td>
</tr>
<tr>
<td>en.wikipedia.org/wiki/Data_Definition_Language</td>
<td>0.415314</td>
<td>9</td>
<td>*</td>
<td>12</td>
</tr>
<tr>
<td>en.wikipedia.org/wiki/Join_(SQL)</td>
<td>0.406125</td>
<td>10</td>
<td>*</td>
<td>17</td>
</tr>
<tr>
<td><a href="http://sqlzoo.net/">http://sqlzoo.net/</a></td>
<td>0.404035</td>
<td>11</td>
<td>*</td>
<td>7</td>
</tr>
<tr>
<td><a href="http://www.stylusstudio.com/sqlxml_tutorial.html">www.stylusstudio.com/sqlxml_tutorial.html</a></td>
<td>0.403724</td>
<td>12</td>
<td>*</td>
<td>11</td>
</tr>
<tr>
<td><a href="http://www.ssmstoolspack.com/Features.aspx">www.ssmstoolspack.com/Features.aspx</a></td>
<td>0.393837</td>
<td>13</td>
<td>*</td>
<td>20</td>
</tr>
<tr>
<td>dev.mysql.com/doc/refman/5.0/en/alter-table.html</td>
<td>0.388677</td>
<td>14</td>
<td>*</td>
<td>13</td>
</tr>
<tr>
<td>en.wikipedia.org/wiki/Data_Control_Language</td>
<td>0.372111</td>
<td>15</td>
<td>*</td>
<td>15</td>
</tr>
<tr>
<td><a href="http://www.postgresql.org/">www.postgresql.org/</a></td>
<td>0.122978</td>
<td>17</td>
<td>*</td>
<td>18</td>
</tr>
<tr>
<td>en.wikipedia.org/wiki/SQL</td>
<td>0.048087</td>
<td>18</td>
<td>*</td>
<td>9</td>
</tr>
<tr>
<td><a href="http://www.geekinterview.com/question_details/12782">http://www.geekinterview.com/question_details/12782</a></td>
<td>not found in SBS</td>
<td>NF</td>
<td>3</td>
<td>NR</td>
</tr>
<tr>
<td><a href="http://www.dmlgroup.in/">http://www.dmlgroup.in/</a></td>
<td>not found in SBS</td>
<td>NF</td>
<td>6</td>
<td>NR</td>
</tr>
<tr>
<td><a href="http://www.wiziq.com/tutorials/data-manipulation-language-dml">http://www.wiziq.com/tutorials/data-manipulation-language-dml</a></td>
<td>not found in SBS</td>
<td>NF</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td><a href="http://www.discogs.com/artist/DML">http://www.discogs.com/artist/DML</a></td>
<td>not found in SBS</td>
<td>NF</td>
<td>10</td>
<td>NR</td>
</tr>
<tr>
<td><a href="http://www.dragon-models.com/">http://www.dragon-models.com/</a></td>
<td>0</td>
<td>NR</td>
<td>5</td>
<td>NR</td>
</tr>
<tr>
<td><a href="http://www.dmlonline.com/">http://www.dmlonline.com/</a></td>
<td>0</td>
<td>NR</td>
<td>8</td>
<td>NR</td>
</tr>
<tr>
<td><a href="http://finance.yahoo.com?q=:=DML.TO">http://finance.yahoo.com?q=:=DML.TO</a></td>
<td>0</td>
<td>NR</td>
<td>9</td>
<td>NR</td>
</tr>
</tbody>
</table>

* : not in top 30 google results
NF: URL not found in SBS
NR: Non relevant to query at all
that were not found by the proposed approach, and

- Ranks 4 to 18 of resources, generated by SSR, were all relevant and were bookmarked in SBS. These were the additional resources that did not exist in at least top 30 results produced by the search engine at the time of evaluation. The graph in Figure 6-20(a) shows the improved results generated by SSR approach consisting of seed resources from the search engine and the filtered resources from SBS ranked according to their relevance to the search query. On comparing results with the originally select resources from the search engine, it is apparent that the crawler downloaded additional web resources which are not seen in the search engine’s result graph line (lying on x-axis). Hence the benefit of using SBS for collecting and ranking relevant resources is also evident.

![Figure 6-17: Description of web documents used by SSR approach for one of the queries](image-url)
Informal e-Mentoring using Multi Agent System

(a) Web resources retrieved from SBS

(b) Rank Comparison of Retrieved Web Resources for a sample query

(c) Frequency Distribution for Graded Relevance Judgment

(d) User Satisfaction (Best Case) – for query term ‘dml’

(e) User Satisfaction (Worst Case)

(f) User Satisfaction (Avg) – for query term ‘sql’
Comparison of SSR with Expert’s Ranking Order: Figure 6-20(b) shows the comparison between the ranks of web resources generated by SSR approach and the ranks given by an expert in the similar field of query topic. One may analyze a reasonable similarity between the two results in the graph with Pearson correlation coefficient of 0.73. This answers the second problem posed in section 6.4.3 i.e., the rank order generated by SSR on crawled web resources is in correlation with the user’s relevance judgment.

Frequency Distribution for Graded Relevance Judgment: Figure 6-20(c) shows the frequency distribution for graded relevance judgment made by the respondents. The non-probabilistic purposive judgment was carried on a four point scale: 3, 2, 1 and 0 representing most relevant, marginally relevant, fairly relevant and irrelevant respectively. In the graph, x-axis represents the resources ranked by SSR. Each bar represents the resources relevance distribution made by 31 respondents for an individual document. It shows that most of the choice has been made between 2 (marginally relevant) and 3(most relevant) by the respondents and 0 (irrelevant) has been at the minimum. The observation clearly shows that most of the retrieved resources are relevant.
Comparing DCG with its Idealist DCG to Determine User Satisfaction: Figure 6-20(d) and Figure 6-20(e) shows the comparison between ranked resources as expected by the learner in idealistic situation (IDCG(n) vector) and ranks as generated by SSR for that learner (DCG(n) vector), for the most satisfied user data and the least satisfied user data respectively. User satisfaction can be considered as inverse of the area between two curves formed by DCG(n) and IDCG(n). More the area between these two curves less will be the user satisfaction. Thus Figure 6-20(d) and Figure 6-20(e) show the upper and lower bound of the user satisfaction for retrieved results. A small area between the curves shows high-quality of user satisfaction, which answers the third question related to the user’s satisfaction. Figure 6-20(d),(e) shows the results of query term ‘dml’ while Figure 6-20(f) and Figure 6-20(g) shows the results for query terms ‘sql’ and ‘group_by’ respectively.

nDCG(n) Vectors Over a Set of Test Collection for Best, Average and Worst Cases: The graph (Figure 6-20(h)) shows the normalized DCG(n) (for n=15) for two respondents’ judgment data with maximum and minimum nDCG(15) measure. The curve with the blue color (upper curve) shows the best result obtained among all judgments which exist in close proximity to the idealistic case. Other one with red color shows the worst case lying below the best case. All other nDCG(n) values lies within these two curves. The curve with green color in Figure 6-20(h) shows avgvect(nDCG) over the set of test collection. The average normalized DCG vector shows the average performance of a particular IR technique over a set of test queries. The curve shows that the proposed approach SSR gradually improves in ranking, although few of the initial ranks have comparatively low gain on an average.

6.5 Discussion

This chapter has discussed an important use of the Concept Ontology and the retrieval, ranking system for disseminating knowledge to a mentee in an informal e-Mentoring system. For the purpose, a system based on the Multi Agent framework was proposed and implemented using JADE and Java programming language. A mentoring task force consisting of primarily five software agents was designed to gather information from the Web, organize it semantically and provide it to a mentee
on his/her request. These software agents were designed with a capacity to create their assistant agents to perform their subtasks and handle multiple mentees without the delay in response time. Their working in different scenarios has been explained through sequence diagrams.

The later part of the chapter discussed the experimental study and evaluation of the complete work which was divided into three phases. The first phase focused on the construction and verification of the Concept Ontology. The content gathering efficiency of various proposed crawlers was studied and evaluated in the second phase. The third phase finally dealt with the evaluation and a detailed analysis of the retrieved web resources’ ranking for the users’ satisfaction.

At large the methods and techniques applied to the system showed a significant performance in gathering web resources from various areas on the Web, viz. the WWW, the Semantic Web and a SBS. The ranking order determined by the system was also in-comparison to the users’ ranking, which showed a reasonable acceptance of the resources recommendation. A system, such as the proposed one would be a real help to an information seeker for the all-time availability of mentors and, to a human mentor who would be benefited by saving enough time for other mentoring tasks than keep searching relevant information for a mentee.