CHAPTER 4

E-LEARNING CONTENT DISCOVERY

In e-learning content management system, discovery of suitable learning materials is an important and challenging task. The requirement of enormous amount of learning materials necessitates an automatic content discovery process. This chapter describes an automatic process of discovering and retrieving appropriate e-learning materials from the web based on an Ontology (ACM classification) and a specially designed learning oriented Topic profile. In this work, IR techniques and technologies are specifically designed to traverse the WWW and collect the educational resources, categorized by topic area. Moreover, Latent Dirichlet Allocation (LDA) model is used to re-rank the learning materials, so that redundancy among the retrieved results is removed.

4.1 INTRODUCTION

The rapid growth of the World Wide Web (WWW) in the last few years has brought enormous opportunities to teachers and learners, in terms of accessing the educational resources. Keyword-based search engines are commonly used for content retrieval today, but there are some problems associated with their use such as, high recall with low precision (retrieval of more irrelevant documents), low or no recall (relevant pages are not retrieved), results that are highly sensitive to vocabulary (Retrieved relevant documents containing terms that are different in meaning with respect to the original query); and results that are single web pages (if we need information
that is spread over various documents, we must manually extract incomplete data from single web pages) (Bianchi et al 2009).

The ability to leverage the vast amount of knowledge and information available on the WWW implies a step in the evolution of an e-learning system (Lawless et al 2008). The challenge of providing the appropriate learning content for formal and informal learners typically depends on the relevancy, topic coverage and quality of content available which in turn depends on the accuracy of the content retrieval system. Moreover, learning materials are typically obtained from an exclusive set of closed corpus content. Steichen et al (2009) stated that the basic problem with this type of sourcing was the need for handcrafted learning objects, enhanced with considerable amount of Meta data. They also discussed the challenge of generating adaptive and personalized hypertext presentation from Open Corpus Content (OCC). The term ‘OCC’ refers to content that is freely available and accessible to the common public and educational organizations. Content can be sourced from web pages, research papers, digital repositories, blogs etc. (Lawless et al 2008).

In order to improve the access methods to educational information, different standards were created like LOM and Dublin Core. SCORM presents some methods for content reusability, but an efficient searching process still remains a major problem (Kardan & Zahmatkesh 2009). OCCS is an existing system that leverages content from Open Corpus Sources for use in e-learning systems, but the educator requires some manual work on the part of content authoring (Lawless et al 2008). Thus, content discovery from the web and content authoring for e-learning is a challenging task in the field of educational data mining. The problems and requirements of an e-learning system can be summarized as follows. i) Discoverability of suitable and relevant content may be difficult. ii) e-learning requires an organization of
content, but the vast quantity of content from WWW is rarely organized and show variants in purpose, topic and format (Lawless 2009). Hence, this work presents content retrieval system for discovering and retrieving relevant learning materials for the purpose of e-learning.

The work described in this chapter, proposed two novel approaches for discovering and retrieving suitable learning materials from the web. One method used ontology, while the other used a specially designed subject oriented topic profile. The main aim of this work is to enable automatic discovery, to assist the instructors, to categorize relevant educational content, and to present high-quality educational materials relevant to a particular course topic without redundancy.

4.2 APPROACHES TO CONTENT DISCOVERY

The existing research work in the area of content discovery and retrieval system for e-learning uses mostly the OCC (WWW, digital content repositories) for retrieving learning materials. Similar work proposed by Brusilovsky & Henze (2007), Lawless (2008) illustrates the various challenges involved in the re-use of materials from WWW for e-learning. Steichen et al (2009) have stated that the problem of exact identification of content from open corpus sources. Here, the system requires a content authoring phase. Annotating the learning material is a time consuming process. However, there is no guarantee that manual content tagging is accurate and objective.

Brusilovsky & Henze (2007) discussed the main problems of using Open content, which is probably not written for the specific purpose (learning need) of the individual learner. Furthermore, Open content is not sequenced in a way which makes learning more directed to the learner’s need, and also not sequenced/linked in an appropriate manner to enable the learning process of
the learner. Identification/Classification of such content against the learner’s need is also difficult. Additional focus to deliver in different styles according to the needs of the learner is required. Lawless et al (2006) discussed the challenges of sourcing, harvesting and incorporating appropriate content for e-learning.

Most of the IR systems focus on storing and viewing documents, approaches for processing queries and determining document relevance, and user interfaces for querying and refining results (Lawless 2009). Focused searching to obtain materials from the web relevant to a particular topic is an active area of research. Khribi et al (2009) used the open source search engine called, “Nutch” in the content modeling phase, to provide automatic recommendations for e-learning personalization. Here, the author improves the educational content indexing phase by adding data (content), meta-data (LOM), and gives specific index fields used in LOM (Learning Object Metadata) to the native Nutch index. However, web information continually changes over time and the basic model representing outdated information can not reflect the user’s interested topics correctly (Gao et al 2010).

Most of the search engines, work on the principle of ranking the pages on WWW. This is one area which has drawn a lot of awareness recently, due to the proliferation of search engines, and many techniques for this purpose have been developed. Balinski et al (2005) have explained document re-ranking methods, to improve the document scores using inter-document relationships. These relationships are determined by distances, and can be obtained from the text, hyperlinks or other information. Many resources which are retrieved from the web are highly redundant. Hence, an appropriate re-ranking of results could help in minimizing redundancy in the retrieved content.
The difficulty in re-ranking of retrieval results is exploring the intrinsic structure of documents in information retrieval (IR). However, one of the problems is that these algorithms handle queries and documents separately. In addition, most of the ranking approaches rely on graph-based techniques, which might ignore several hidden information in the retrieval results (Zhou & Wade 2009). However, LDA-based content retrieval is a promising technique for ranking the whole corpus, to identify the internal structure, and to find the latent structure of “topics” or “concepts” in the initial retrieval results. We discuss the two proposed approaches in the following sections. The first approach includes ontology based query method, concept focused crawler and concept-term based ranking. In open and distance learning, ontologies are used as knowledge bases for e-learning enhancements, searching systems, educational recommenders, and question answering systems which support learners with much needed resources (Nguyen & Yang 2012). Ontology is a formal, explicit specification of a shared conceptualization (Noy & McGuinness 2001). In education and e-learning, many researchers have built learning systems, which takes the benefits of ontologies. Li & Rui (2005) proposed a way to organize and compose learning objects with the use of ontology for a recommendation mechanism. Ana et al (2009) presented a domain ontological model to build a recommender system in which to provide a decision making facility. Jagadish et al (2011) presented domain ontology based re-querying approach for query expansion to discover content from open corpus sources. The authors discussed that incorporating domain knowledge in the form of ontology can help to perform sense-disambiguation in the retrieved content. Saman et al (2012) developed a knowledge-based and personalized e-learning system based on ontology. These ontologies were constructed manually with the help of expert knowledge obtained from many resources and documents. In e-learning environments, usage of ontology aims to provide mechanisms to enhance the process of searching and discovering learning materials. In
addition, it provides the facility to organize and display information, which makes it easier for learners to draw connections, for instance, by visualizing relationships among concepts and ideas (Ibrahim 2012).

4.3 ONTOLOGY BASED CONTENT DISCOVERY

Most existing content discovery techniques rely on indexing keywords. Unfortunately, keywords or index terms alone cannot adequately capture the document contents, resulting in poor retrieval performance (Lawless et al 2009). Typically, the information need is expressed as a combination of keywords. However, in this work, we use topic learning terms associated with topic under consideration extracted from the domain ontology. These topics and learning terms are used in the concept based query method which is described in this chapter. In addition, this work proposes a concept and term based ranking system for ordering the documents from search engine. Sophisticated search engines can handle e-learning materials easily. However, the retrieved resources may not have a complete coverage of topics which the instructor actually requires for content authoring. Moreover, a number of resources which are retrieved are highly redundant (Brusilovsky & Henze 2007). Hence, appropriate ranking of documents using concept and topic learning terms possibly will help in retrieving topic related documents and reducing redundancy of the retrieved content. Here, the ranking system exploits the concept-document similarity of the document collection. These ranked documents could then be used as seed documents for our crawling system. A focused crawler can be defined as a web crawler, which actively seeks, acquires indexes and maintains pages on a specific topic, which represent a relatively narrow segment of the WWW. The goal of a concept-focused web crawler is to selectively seek out pages that are relevant to a predefined set of concepts associated with a topic. Besides sourcing web pages based on its content, concept-focused crawling allows a web crawler to
process specific sites to greater depths than general purpose crawlers. Focused
 crawlers can spend more time perusing highly relevant sites rather than
 attempting to attain broad coverage of the entire WWW in a breadth-first
 manner (Chakrabarti et al 2002). After crawling, concept-terms similarity
 ranking system has been used to rank the crawled documents. As a result,
 highly relevant pages can be discovered that may be overlooked by normal
 focused crawlers, at the same time filtering redundant pages to avoid
 unnecessary paths.

Figure 4.1  Retrieval of educational content from web based on domain
 ontology

Figure 4.1 shows that the system for retrieval of educational
 contents from the web for e-learning consists of 3 phases. The first phase is
 the Ontology based query method for obtaining seed documents. The queries
 are generated based on the ontology concepts and Term Frequency- Inverse
 Document Frequency (TF-IDF) based scoring mechanism is used for query
expansion. The output of this phase is set of documents, which is used as seed documents for crawling system. The second phase is the Concept-term based ranking of seed and crawled documents. This phase takes a set of documents and ranks them according to concepts and terms associated with these concepts. The final phase is the Concept-focused crawling system which takes a set of ranked seed documents and produces a set of crawled documents. These documents are again ranked by using concept-term based ranking which gives final set of crawled documents.

4.3.1 Ontology based Query Method

While dealing with content retrieval from the web for educational contents, using the TF-IDF for determining terms for query expansion is inadequate and different methods are required to find out related concepts. In addition, a simple TF-IDF based re-querying scheme can leave out some potentially relevant sub concepts related with the query term (Jones 2004). An existing approach for finding the educational content from the web is determined by a set of keywords in a file (Lawless 2009). However, such an approach which is completely TF-IDF based would not be able to identify related words. Similar to the work described by Luong et al (2009), the proposed system also used the concepts of the ontology to query the web to obtain seed documents. Compared to the existing work described here, this work used an ontology specially designed for the computer science based concepts, which is based on the ACM classification hierarchy (ACM Classification 2012). The association of terms to concepts for specific purposes has been used by InfoWeb (Gentili et al 2003) a filtering system using user profiles in a digital library scenario. Here the semantic network used to represent the user profile has nodes representing concepts and as more information is gathered about the user the profile is enhanced by associating additional weighted keywords with these concept nodes. The idea in this work differs in considering the set of topic learning terms typically designed for a
particular topic and additionally having concepts from the ACM classification. In order to provide a rich content exposure to the user, this work takes TF-IDF score and ontology based query expansion specifically designed for e-learning. The search using concepts and topic learning terms from the ontology retrieves a set of seed documents. The ontology based query method is required, because it integrates some additional information about relationships. For example, when the query is given as input to the search engine for the learning materials relevant to ‘Network transmission media’, the keyword based search (Lawless 2009) lacks to identify “Guided media, unguided media” as related words, because it is unaware of the fact that a part-of relationship exists between ‘Guided or unguided media’ and ‘Network transmission media’. In the next section, we discuss the ranking scheme, which is used to rank the seed and crawled documents.

4.3.2 Concept-Topic Learning with Terms based Ranking

The majority of the resources retrieved from the web may contain repeated information in different forms (Brusilovsky & Henze 2007). This redundant information affects the learners’ interest and the efficiency of the system. Identifying the most relevant results to a query is a central problem in web search; hence ranking of documents has received a lot of attention. An existing Ontology based system for ranking documents, combines conceptual, statistical and linguistic features of documents, and also expanded the query with its related concepts before comparing the documents (Shamsfard et al 2006). However, our concept-term ranking system is oriented towards the retrieval of e-learning content and is based on two basic principles. i. Documents having concepts of the topic along with learning terms, and sub concepts associated with these concepts are better resources for e-learning, since learning materials should have high coverage of topics, to ensure content quality. ii. Documents which are about sub concepts of terms and not covered in the documents hitherto ranked are ranked as better resources, since
we want to avoid redundancy of materials and also ensure better coverage of
the topics. The ranking mechanism does four comparisons based on concepts
and topic learning namely, the number of concepts from the ontology that the
seed document has, the similarity of concepts with concepts in other seed
documents, number of topic learning terms associated with concepts and
coverage of topic learning terms of the concept’s and sub-concepts. If the
document contains more topic learning terms, then it will receive a higher
rank. Initially, sub-concept terms (inner most level) are considered for
ranking. Then the main concept, sub-concept’s name and the length of the
document are considered. The document specific features such as the length
of the document, number of concepts, total topic learning terms, and
individual frequency of terms have been calculated for each document.

\[ D_w = \sum_{i=1}^{n} TF_i + Term_{i, level} \] (4.1)

In Equation (4.1), ‘\( D_w \)’ refers to the weight of a single document,
‘\( TF_i \)’ refers to the number of occurrences of the \( i^{th} \) term, ‘\( Term_{i, level} \)’ refers to
the ontology level weight for the \( i^{th} \) term. Here, weight of the \( Term_{i, level} \) will
be given based on the ontology level. We consider three levels in our domain
ontology. If the \( i^{th} \) term presents in inner most level of the ontology then the
weight will be assigned as ‘1’. For sub-concepts (2\(^{nd}\) level), the weight will be
assigned as ‘0.5’ and finally concepts obtained (1\(^{st}\) level) the weight as ‘0.25’.
Normally, the documents which contain general concepts (1\(^{st}\) level) can be
identified easily from the web. From e-learning perspective, the inner most
level of ontology terms is important while creating learning content, when
compared to other levels. Therefore, we assigned higher values for inner
levels.

After obtaining the seed documents, the system perform the ranking
mechanism to select and crawl resources that are specifically suited to e-
learning. This type of ranking ensures the coverage of topic and concept. The
ranking mechanism has been used to rank the seed documents from search engine and also rank the crawled documents for filtering out appropriate documents for e-learning. The concept based focused crawling is explained in the following section.

4.3.3 Concept-Focused Crawling System

Focused crawler for intelligent web based content management system mined the educational contents from university websites in the form of course pages (Sajjanhar 2008). This system only mined from the university websites using, Hidden Markov Models (HMMs) and Conditional Random Fields (CRFs). Li et al (2005) proposed a method using a decision tree on anchor texts of hyperlinks for focused crawling. They have used feature selection for picking keywords for decision trees. Ehrig & Maedche (2003) proposed an approach for document discovery building on a comprehensive framework for ontology-focused crawling of web documents. This framework included a definition of the knowledge structure and its lexical presentment, relevance computation strategies, and an empirical evaluation of the overall framework. Another framework ontology learning process includes, crawling, classifying and extracting relevant information from documents. SVM classification has been used here to filter out documents in the searched results (Luong et al 2009). However, our crawling system uses the concept-term ranking to filter out the searched results. Most of the above focused crawlers have used keywords alone for focused crawling system. However, our crawling system initially ranks seed documents so as to crawl and obtain documents suitable for e-learning. We rely on background knowledge with concepts and associated topic learning terms, which are compared with the contents of the crawled documents. The concept-term ranking had been used again to obtain the final set of crawled documents, which is needed to avoid redundancy of content.
4.3.4 Experimental Results

This section reports results from tests conducted to estimate the influence of the suggested ranking method. The experiment uses documents initially produced by the Google search engine. In this work, relevant documents are required for creating learning content. Redundancy will increase, if large numbers of documents are considered which in turn will increase the execution time. The relevancy has been checked based on the match process with the domain ontology concepts related to the terms present in the learning material.

Table 4.1 shows the sample query results of documents produced by the Google search engine and relevant documents after applying concept-topic learning terms based ranking. In Google search, the number of documents retrieved is more. However, we are getting lesser amount of domain ontology based relevant documents.

<table>
<thead>
<tr>
<th>Query</th>
<th>No. of documents (Google Search)</th>
<th>Relevant documents</th>
<th>Relevant documents after ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network topology and its subtopic</td>
<td>200</td>
<td>180</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>270</td>
<td>180</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>360</td>
<td>230</td>
</tr>
</tbody>
</table>

Table 4.2 shows the Precision value before and after ranking. Here, the Precision value before ranking decreases, since the topic relevant documents of learning materials are fewer compared with Google searched documents. After ranking the precision value increases and more accurate learning materials relevant to domain ontology are highly ranked. In addition, this ranking method removes the redundant documents from the topic relevant
documents. The precision value is high for first 200 documents, since the documents retrieved from the Google initially, is highly relevant to the querying topic.

**Table 4.2 Precision results before and after applying ranking method**

<table>
<thead>
<tr>
<th>No. of documents (Google Search)</th>
<th>Precision before ranking</th>
<th>Precision after ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>0.9</td>
<td>0.72</td>
</tr>
<tr>
<td>500</td>
<td>0.54</td>
<td>0.66</td>
</tr>
<tr>
<td>1000</td>
<td>0.36</td>
<td>0.63</td>
</tr>
</tbody>
</table>

**4.3.5 Topic Specific Evaluation**

Our ontology based content retrieval module retrieved documents for the following topics: Communication, Computer Network Fundamentals, Computer Network Services, Advanced Computer Networks, and Networks Security. Table 4.3 shows the Main topics, number of Google searched documents and number of resulted documents after ranking. The total Documents retrieved are 2600, the total documents relevant to the query without redundancy are 684 and the precision value is 0.263.

**Table 4.3 Results for particular topics**

<table>
<thead>
<tr>
<th>Topic name</th>
<th>No. of documents from GSE*</th>
<th>Filtered Documents</th>
<th>Documents after ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication</td>
<td>550</td>
<td>238</td>
<td>130</td>
</tr>
<tr>
<td>Computer Network Fundamentals</td>
<td>500</td>
<td>287</td>
<td>147</td>
</tr>
<tr>
<td>Computer Network Services</td>
<td>600</td>
<td>324</td>
<td>172</td>
</tr>
<tr>
<td>Advanced Computer Networks</td>
<td>500</td>
<td>351</td>
<td>122</td>
</tr>
<tr>
<td>Networks Security</td>
<td>450</td>
<td>233</td>
<td>113</td>
</tr>
</tbody>
</table>

GSE* - Google Search Engine
While using domain ontology for content retrieval, we can have concepts, sub-concepts, topic learning terms for query generation and ranking. However, we could not have the pedagogy details of e-learning when using domain ontology. Therefore, we design a topic profile especially for e-learning usage, which contains chapters, topics, annotated terms that incorporate the details from more than one curriculum, and pedagogy information from the learning viewpoint. This topic profile based second approach (will be discussed in the following section) includes an automatic generation of simple and concatenated search engine queries, topic focused crawler and LDA based re-ranking scheme for prioritizing the search results, all based on a specially generated topic profile.

4.4 TOPIC PROFILE BASED CONTENT DISCOVERY

This work presents an automatic generation of search engine queries to obtain seed URLs for a topic focused crawler. Normally, hyperlinks presented in the web pages are crawled in a breadth first manner. However, educational contents obtained through crawling need to focus on the content suitable for e-learning. As we discussed earlier in section 4.2.2 the retrieval of redundant information from the web is a serious problem to be tackled, as it reduces the efficiency of the system. Using topic profile terms and levels queries are generated for specific topics. Furthermore, the topic focused crawler is used to avoid crawling of hyperlinks which are not relevant to a specific topic and crawls considers only the web pages which are suitable for our e-learning content management system. In this work, we introduce a new idea of a ‘topic profile’ (profile of the subject in contents of chapters, topics, and annotation terms represented as a spanning tree) which is specially designed for the Content Retrieval System. In addition, we propose a topic profile based LDA for re-ranking scheme to prioritize the search results so as to maximize the topic exposure and minimize the redundancy among the
retrieved contents. We have constructed the topic profile as a spanning tree for a particular course (subject), incorporating details from more than one curriculum, which results in a hierarchical representation of the topics, subtopics and terms that should be covered in a topic from a learning view point. Topic profile is represented as a hierarchical illustration of keywords for a particular topic. Here, the topic profile is used for generating queries, topic focused crawling and also for re-ranking of the retrieved results. Figure 4.2 shows the Architecture diagram for Content Retrieval and Re-ranking system.

4.4.1 Topic Profiling

Generally, a content retrieval system uses domain ontology or set of key words or concepts for searching and retrieving documents from the web.
Here, we have constructed a topic profile to search and discover the documents. The input to the topic profiling module is a collection of documents (which can be text books, presentations or references). Here, the ‘table of contents’ page of all documents are extracted. Each topic in the index is represented as a node in a graph G. The contents corresponding to each topic in the index are clustered together, and the Kullback Leibler (KL) divergence (Kullback & Leibler 1951) between each index’s cluster probabilities is estimated. Given two topic cluster probabilities, \( P(i) \) and \( Q(i) \) the KL Divergence is given by

\[
(P||Q) = \sum P(i) \ln \frac{P(i)}{Q(i)}
\]

This KL Divergence forms the cost of the edge connecting the nodes in the graph G. The minimum spanning tree of the Graph G gives the order in which the topics can be grouped. If the cost between the nodes is lesser than the threshold, they can be merged as a single node. This tree is the topic profile that needs to be followed in the course (Profile of subject). The reason for constructing a minimum spanning tree is

- A tree is acyclic; so, a topic discussed once will not be revisited.
- A minimum spanning at each point finds local optima, and jumps to that node which helps us to create a hierarchical illustration of keywords for a particular topic.

Figure 4.3 shows the flow chart of the topic profiling scheme. Figure 4.4 shows the sample spanning tree for the concept ‘Process’. The nodes 1, 2, …, 12 are ‘Process concept’, ‘Process scheduling’, ‘Inter Process Communication’, ‘Threads’, ‘Multithreading models’, ‘CPU scheduling’, ‘Scheduling algorithm’, ‘Thread scheduling’, ‘Process synchronization’,
‘Critical section Problem’, ‘Semaphores’, and ‘Deadlocks’ respectively. This topic profile shows that topics are not revisited, and transition at any point is made only to related concepts. For example, the ‘Critical Section Problem’ occurs in a hierarchy only beneath the Process synchronization and Process concept. Similarly, ‘Thread scheduling’ is visited only after visiting ‘Threads’ and ‘Multithreading models’ respectively.

Figure 4.3  Flowchart of topic profiling
4.4.2 Automatic Query Generation

From the perspective of e-learning, it is critical to retrieve documents that cover the essential concepts of a specific topic based on keywords. As, we discussed already in section 4.2.1 most of the systems do not consider any structure and relations among the topics for querying educational content. Hence, we present an automatic generation of simple and concatenated queries based on topic profile levels and terms. This system incorporates the additional knowledge about the topics, in terms of topic words, and levels and thus establishes the connection between the main topic and sub topic on a level by level basis.

For example, to generate queries for Operating System concepts, examples of simple queries include Operating System, Process Management, Memory Management, Paging, Semaphore, etc… Examples of concatenated queries include Operating System + Process Management, Operating System + Process Management + Threads, Operating System + Process Management
While generating concatenated queries, we have used topics, terms, and levels from the topic profile. This topic profile based simple and concatenated queries, are then given to a standard search engine, to obtain URLs which are relevant to the specific topics that we have used in the queries. These relevant URLs specific to the query terms are then used as seed URLs for our topic focused crawler. This topic profile generated using simple and concatenated queries provide a good set of seed URLs for our focused crawling system.

Algorithm 1 describes the processes of content retrieval from the web; here, we have mentioned the input, output and process of content retrieval. Part (a) explains the steps of automatic query generation. Once we have constructed the topic profile, the automatic query generation system reads the topic terms and levels from the topic profile. The system generates a simple query for all topic terms which is presented in the topic profile separately. After that, for each simple query, concatenated queries have been generated by adding the topic terms one by one, if the terms are present in the inner levels of the topic profile. These simple and concatenated queries retrieve documents from the web through the search engine. These documents are then used as seed documents for our topic focused crawler. Also, the redundant links are removed while extracting seed document URLs.

**Algorithm 1:** (Content Retrieval from Web)

**Input:** A topic profile document tp_doc, seed document as SD

**Output:** Document Set-A corpus of crawled documents of all topics associated with tp_sq (simple query), tp_cq (concatenated query) based on tp_doc
Process:

a) For all topics in tp_doc

\[ tp_{sq} = \text{topic (i)} \]
\[ tp_{level} = \text{level of topic (i)} \]
\[ tp_{cq} = tp_{sq} (i) + tp_{cq} (i) \]

For all \( tp_{sq}, tp_{cq} \)

\[
\text{search_links.add (top 10 search results obtained from web)}
\]

Remove redundant data from search_links

End for

b) While (all the crawling levels)

Do

For each SD (i) in Corpus

\[ \text{all_links} = \text{all hyperlinks parsed from SD (i)} \]

For each link (i) in all_links

If (link_word(i) == each tp_sq(i))

Compare link (i) in selected links

<Document Set>corpus. Add (web page of the URL in link (i))

End for

End for

End do
4.4.3 Topic Focused Crawling

Web crawlers play an important role in searching the web content from the WWW. They crawl through all hyperlinks present in the web pages continuously. These crawlers are like public libraries trying to cater to everyone, and do not specialize in particular areas (Farag et al 2012). E-Learning users are increasingly feeling the need for highly specialized and filtered learning materials, where they can explore their interest in depth. Hence, we propose a goal-directed crawling as a powerful mechanism for topical documents discovery. The focused crawler is a system that learns the specialization from examples, and then searches the web, guided by relevance, and obtains high quality of web pages.

Even though focused crawling and classic crawling starts from the same origin set, focused crawling obtains relevant pages gradually, but the classic crawling quickly loses its way. Focused crawling is also capable of searching out and discovering valuable resources that have more links away from the starting set, thereby cautiously reducing thousands of pages that may lie in this same circle (Chakrabarti et al 1999). A topic profile based crawling is needed, as it includes a number of additional topic information like terms, levels and connections between topics. We state that such connections play a key role in an e-learning scenario, where a learner is expected to obtain more knowledge by incorporating all the concepts in the topic of his interest.

In this work, the topic focused crawler has been designed to obtain seed URLs from search engines using simple and the concatenated query system and crawl the hyperlinks which are matched with the topic terms present in topic profile. The main step in the crawling activity is to find the links in a web page. All the anchor tags in the downloaded HTML pages are retrieved and parsed using the HTML parser. In order to do a topic focused
crawling, the link texts have been checked and compared with the topic profile. Those links which have their link text in the topic profile alone have been selected and crawled. While crawling the web, the redundant links are eliminated even before obtaining the web pages of those links. Finally, the crawled documents are correlated with the topic profile to provide the best classified educational materials based on topics.

Algorithm 1 provided in the part (b) explains the steps of Topic focused crawling to obtain the relevant documents from the web. Initially, all the hyperlinks from each seed document have been parsed. Then these hyperlinks need to be compared with the topic profile terms file. If the links matches, the crawler allows the links to processed; otherwise, it omits the hyperlinks.

4.4.4 Re-ranking of the Learning Materials

In E-learning, redundancy in the contents may demotivate the learner, as the learner may find the materials to be inappropriate to his needs. This may lead the learner to skip the topic and move on to other topics. Current search engines have designed their ranking schemes based on pages, with the search terms appearing in the HTML title tag being frequently assumed to be more appropriate to the particular topic. Search engines also check if the search keywords come close to the top of a web page, such as in the headlines or in the first few paragraphs of the text. The intention of the re-ranking phase, is to rearrange a set of documents in order to the improve retrieval accuracy at the top ranks of the final retrieval set. Most of the approaches signify the document entities as a linked graph G. It is generally built by links inferred from the content information as a nearest-neighbor graph which may ignore several hidden topics within the retrieval set (Zhou & Wade 2009). The topics which are not detailed while describing some other
key topics are said to be hidden topics. Hidden topics are mostly used to classify the documents. For example, while describing the topic memory management in Operating System; the hidden topics are threads, segmentation, paging, kernel etc.…

The existing approaches to the content retrieval for e-learning, do not handle the redundancy while retrieving educational materials (Bianchi et al 2009; Lawless et al 2008). Our re-ranking system minimizes the redundancy among retrieved documents. The re-ranking system uses topic profile based LDA which exploits the hidden structure of the documents with respect to specific queries. Relatively depending on graph based methods to recognize the internal structure, the approach attempts to discover the latent structure of “topics” or “concepts” in the initial retrieval results (Zhou & Wade 2009). LDA is capable of consistently bringing significant progress over various baseline systems that are used to improve the retrieval accuracy at the very top most ranks of the final document results. LDA helps to overcome the problem of topic drift (Zhou & Wade 2010). Some documents start with a specified topic, and move on to various topics that are irrelevant to the current topic. This topic drift can be avoided using LDA, as it helps to find various topic mixtures for each document.

The work described in this chapter, uses Plate Notation LDA model which incorporates topic profile words as vocabulary. Using plate notation, the dependencies among the many variables can be captured concisely. The rectangles are “plates” representing replicates. The outer plate denotes documents and the internal plate represents the repeated option of topics and words within a document. M represents the number of documents; N is the number of words in a document. Figure 4.5 shows the plate notation for the topic profile based LDA model.
Figure 4.5 Plate notation for topic profile based LDA model

\[ \alpha \] is the parameter of the Dirichlet prior to the per-document topic distribution.

\[ \beta \] is the parameter of the Dirichlet prior to the per-topic word distribution.

\[ \theta_i \] is the topic distribution for document \( i \),

\[ \varphi_k \] is the word distribution for topic \( k \),

\[ z_{ij} \] is the topic for the \( j^{th} \) word in the document \( i \), and

\[ w_{ij} \] is the specific word.

tpword is the Topic profile vocabulary.

The \( w_{ij} \) are the only observable variables, and the other variables are latent ones (Blei et al 2003).

In the re-ranking setting, we estimate that the probability of a document \( d \) generates \( w \), using a topic profile based mixture model LDA. It uses a convex combination of a set of component distributions to model observations.

\[
LDA_d(w) = \sum_{z=1}^{k} p(w|z)p(z|d)
\] (4.3)
In Equation (4.3), a word \( w \) is generated from a convex combination of some hidden topics \( z \) and Topic profile vocabulary \( t \) \( \text{word} \), where each mixture model is a multinomial distribution over terms that correspond to one of the latent topics \( z \).

\[
\text{LDA}_d(w_1, w_2, \ldots, w_n) \equiv \prod_{j=1}^{n} \text{LDA}_d(w_j)
\]

(4.4)

Algorithm 2 describes the steps of Re-ranking of the documents. Here, we have explained the input, output and process of the Re-ranking system. For all the crawled documents in the preprocessed corpus the topic mixture and word frequencies has been calculated using the topic profile based LDA. Then, the content similarity of all documents has been measured to remove redundant ones from the corpus. Finally, the ranking process has been done according to the topic profile level, length of the document, weight of each word, and topic mixture of the word.

**Algorithm 2: (Re-ranking)**

**Input:** Crawled document corpus as Document Set, LDA model, preprocessed_corpus PD, topic profile as \( t \)

**Output:** Prioritized document collection with non-redundant

**Process:**

For each PD \((i)\) in preprocessed _corpus_

Using LDA calculates topic mixture and word frequencies

Sort the calculated word frequency

If word frequencies of each word and length of any two or more PD \((j)\) matches

preprocessed_corpus. Remove (PD \((j)\))

End if
End for

For all remaining PD (i) in preprocessed_corpus

Rank according to tp_level, document length, weight=word frequency/document length, High frequency words using topic mixture

4.4.5 Experimental Description

Our Topic Profile based approach to automatic query generation and the Re-ranking algorithm were both tested, through the retrieval of 15,210 documents from the WWW. Initially, we constructed a topic profile based LDA model, using the JgibbLDA package, from a training set of 15210 documents retrieved from the web. The following Figures 4.6(a) to (c) show a sample snapshot of the word-topic probabilities for topic 1, topic 2, and topic 3 respectively from the topic profile based LDA model:

![Figure 4.6 Word-topic probabilities for topic 1, topic 2, and topic 3](image)

The measures topic coverage and non-redundancy are used to show the effectiveness of our approaches of Content Retrieval and Re-ranking system.
1. Topic coverage is the fraction of the number of topics covered in all the documents and the total number of topics presented in the topic profile.

2. Non-redundancy is defined to calculate the cosine similarity values of the set of documents retrieved from the web. If each document has different cosine values, then it will be considered as a non-redundant document. We consider this to be an excellent measure to find the dissimilarity of content in the document corpus.

**Content redundancy elimination**

While extracting links for topic focused crawling, duplicate links have been eliminated even before obtaining the web pages of those links thereby aiding to reduce the content duplication to some extent. On the other hand it is also possible for two or more saved web pages to have the same content inspite of having different URLs. For example, two different links can point to two different sub topics of the same web page, thus leading to content redundancy. In order to eliminate such type of content redundancy, we check the document length and count of the topic profile words. If two web pages contain the same count of the topic profile words and the same document length, then any one of the documents can be removed from the resulting corpus.

**4.4.6 Results**

To verify the effectiveness of our automatic query generation system, the topic coverage of the simple and concatenated query results are measured. Topic coverage is used to find out the relevance of the documents
obtained from the search engine queries. In Figure 4.7, we show the topic coverage of the simple and concatenated query results. For a simple query, the relevant documents are 11045, and for a concatenated query, the relevant documents are 12968 around the retrieval of 15210 documents. Since the concatenated query (Operating System + Process Management + Threads + Linux Threads) retrieves with high topic coverage, it searched for documents on the web in a more focused manner. In addition, we have compared the normal web crawler with our topic focused crawler to show the efficiency of retrieval. Normal web crawler is crawling through all hyperlinks presented in the web pages in a breadth first manner continuously. Since the topic focused crawler omits the unwanted crawl of hyperlinks, the processing time is reduced. This is measured by calculating the number of crawls for a set of seed documents.

![Topic Coverage Diagram](image)

**Figure 4.7  Topic coverage of query results**

Figure 4.8 shows the number of crawls of the normal web crawler and the topic focused crawler. We have also measured the topic coverage to obtain the relevance of the crawled documents. For a set of seed documents,
the number of crawls of the normal web crawler is 27345, and for a topic focused crawler it is 15210. Since the topic focused crawler neglects the unnecessary crawl of hyperlinks, the number of crawls is reduced and the topic coverage is increased.

![Number of crawls](image)

**Figure 4.8 Comparison of Normal web crawler and Topic focused crawler**

Figure 4.9 shows the degree of improvement of the topic coverage of the crawled document collection by using the focused crawler. Hence, the topic focused crawler documents shows high precision. Furthermore, we evaluated the re-ranking scheme to show the improvement of non-redundancy among the results. Content similarity has been measured to remove the redundant documents. Non-redundancy will be evaluated by measuring the content similarity and topic coverage of the documents, before and after re-ranking.
In Figure 4.10 the graph indicates the scale of improvement on the measure of diversity of the content that we retrieve and accumulate in the document corpus. By using the Re-ranking, the enhancement to the degree of non-redundancy of content is increased by 18%.

Figure 4.9 Topic coverage of crawler results

Figure 4.10 Comparison of non-redundancy
4.5 SUMMARY

In this chapter, we have proposed two approaches for discovery and retrieval of educational materials, from the web based on Ontology and Topic profile. We have shown that the topic focused crawler retrieves more relevant documents, when compared with the normal thereby improving the topic coverage and reduce the number of unwanted crawls. We have also proposed re-ranking scheme for prioritizing the results, which are obtained from the search engine. The re-ranking system also minimizes the redundancy among the results, so as to improve the diversity of the retrieved content. Also, we have evaluated our proposed approaches using the measures namely, topic coverage, non-redundancy, content similarity and relevance based metrics. The results obtained from content discovery component was used to create learning objects and those learning objects are optimized and sequenced which is discussed in chapter 6. The next chapter provides the details of learner profiling and classification based on the learners’ behavior.