CHAPTER 1

INTRODUCTION TO e-LEARNING AND INFORMATION RETRIEVAL

1.1 e-LEARNING

Intentional use of electronic media and Information and Communication Technologies (ICT) in teaching and learning process (Naidu 2006) is referred to as e-learning, where “e” denotes “electronic”. It can also be described by many other terms including online learning, virtual learning, distributed learning, network and web based learning. e-learning includes all educational activities carried out by individuals/groups working online/offline and synchronously/asynchronously through network/standalone computers and electronic devices.

Individualized self-paced e-learning - online refers to situations where individual learners access learning resources like database or course content online through Intranet/Internet. Individualized self-paced e-learning - offline is about a learner using learning resources like database/computer-assisted learning packages.

Group-based e-learning synchronously means situations where learner groups work together in real time through Intranet/Internet. Group-based e-learning asynchronously means situations where learners work over an Intranet/Internet with participants exchanges occur with a time delay.

e-learning (Marković 2010) enables higher interactivity among professors and students and study material coverage in both undergraduate/graduate students. Further, professors and assistants ensure that students' critical thinking is developed, and to provide them freedom in discussion, topics choice, exchange of ideas and information, and expansion of knowledge.

As the development of technology grows, e-learning helps students in their studies in an easy manner, anytime and anywhere. e-Learning has become a popular and acceptable way to study due to its flexibility and better innovativeness regarding introduction of new/contemporary programs as compared to traditional faculty.

Also, many faculty who opted for e-learning started implementing various software packages supporting online learning in addition to application of different studying modalities.

1.2 TEN PRINCIPLES FOR SUCCESSFUL e-LEARNING

Principle 1: Match to the Curriculum (Anderson, et al., 2005): Pedagogy should match and be aligned with appropriate curriculum via clear objectives; content relevance; student activity appropriateness and nature of the assessment.

Principle 2: Inclusion: Pedagogy should include practice seen regarding different types/range of achievement and physical disabilities which can be specifically supported by e-learning.

Principle 3: Learner Engagement: Learners should be engaged and motivated by pedagogy.

Principle 4: Innovative Approaches: The reason of learning technologies rather than non-technological approach leadings to similar end used should be evident. e-learning should suit specific purposes.

Principle 5: Effective Learning: This can be demonstrated through many ways; using various approaches in the learning platform to permit a student to choose what suits him/her.

Principle 6: Formative Assessment: Pedagogy should ensure formative assessment.

Principle 7: Summative Assessment: This must be valid and reliable, and handle various achievement levels; it should also be free from learner's adverse emotional impact.

Principle 8: Coherence, Consistency & Transparency: Student activity and assessment should match each other and pedagogy should be internally coherent/consistent in matching objectives and content.

Principle 9: Ease of Use: e-learning should ensure ease of use and transparency.

Principle 10: Cost-Effectiveness: Technology solutions should be justifiable /affordable with sustainable costs.

1.3 STRATEGIES FOR e-LEARNING

(i) The e-learning must be participant centered.

(ii) A case must be made when value is not obvious and when point to data needs assessment. Presentation of a problem/case to participants improves clarity.

(iii) The program must ensure opportunities for success and not failure/uncertainty. To motivate/maintain involvement participants should nurture self-efficacy.

(iv) Make it real to ensure that programs should match audience in both topic/level.

(v) As e-learning relies on involvement/generosity, reveal what participation will result in.

(vi) Make it active/thought-provoking: A virtual coach reveals choices pointing out missed opportunities.

(vii) Make it human: Showcase people/emotions/successes. Reveal how people feel about what can be learnt /achieved.

(viii) Guide/track participants: Controlled experiments indicate that when novel information is dealt with, learners must be taught what and how to do.

(ix) Situate e-learning within a blend: A blended experience transcends a single experience scheduled for a specific time/place.

(x) Relationships, collaboration and teaming should be part of effort as the idea of an online community is now increasingly important.

(xi) Make it WOW! Which is when everything comes together to generate something dramatic, compelling, valued, and authentic. Something that attracts participants and involves them.

(xii) Measure and continuously improve e-learning and learning management systems (LMSs) ensure executives are comfortable with technology-based information on compliance/risk avoidance (Allison, et al., 2008).

When forming roles and responsibilities within learning systems, current and future directions should first be identified. This is done by first considering traditional learning actors roles. Two important categories of e-learning (Simic et al 2011) are experiential (significant) learning, and cognitive (meaningless) learning. Many methods contribute to effective knowledge building, but many also keep projects/problem-based learning as the cynosure. Problem solving techniques called problem-based learning can engage learners in knowledge building actively.

In addition to problem and project-based learning, similar learning methods including active learning, inquiry-based learning and service learning exist. As regards active learning, to ensure active involvement in a learning process, students should perform analysis, synthesis and evaluation, which means that listening alone is not enough. Active learning requires active part in comprehension by discussing, writing, playing simulation game roles and problem solving for learners. Inquiry-based learning recognizes that science topics are question-driven and open-ended to understand which, learners have to learn how to pose questions, perform investigations and obtain results from this basic aspect of science.

Based on software communication characteristics and resources for e-learning, three different e-learning environments are distinguished:

(i) Self-study,

(ii) Asynchronous, and

(iii) Synchronous.

1.4 ADVANTAGES AND DISADVANTAGES OF e-LEARNING

e-learning applications/processes (Anand et al 2012) include computer-based, web-based and technology based learning, in addition to virtual education opportunities. Content delivery is through internet/ intranet/ extranet and audio or video tape, satellite TV, and CD-ROM including media as text, image, animation and video and audio streaming.

e-learning's main attribute is more to access information/resources. This refers to the access of information/resources any time, any place or any pace based on one's convenience. Another characteristic is access of multimedia based resources. They are various media types like text, audio, video, animation, graphics, picture in network and communication technology are supported, and which ensure information access by not only text/pictures but also through supported animations, videos, presentations and audio.

Currently e-learning is a highly emerging knowledge tool, providing a method to deliver knowledgeable contents through CD, DVD, multimedia and other tools. Its main drawback is availability of bandwidth, e-learners willingness, and skill sets to deliver material to learners.

For most regions, e-learning did not just open up "existing learning structures/content to new customers". Many regions emphasized e-learning's new methodological potential to "transform learning process", its advantages being its greater interactivity, connectivity, adaptability, and capacity to promote digital and key skills.

1.4.1 Advantages of e-learning to the Trainer or Organization

The advantages of e-learning include

(i) ***Reduced overall cost*** is the major factor in adopting e-learning. Reduced time away from the job may be it's a positive offshoot.

(ii) ***Consistent content delivery*** is possible through asynchronous, self- paced e-learning.

(iii) ***Expert knowledge*** is communicated and also captured through e- learning and knowledge management systems.

(iv) ***Proof of completion and certification,*** which are major training initiative elements, can be automated.

(v) ***On-demand availability*** ensures that students complete training during off-hours/from home.

(vi) ***Self-pacing*** for slow/quick learners increases satisfaction and lowers stress.

1.4.2 Disadvantages to the Trainer or Organization

(i) ***Up-front investment*** for an e-learning solution is high because of development costs. Budgets/cash flows should be negotiated.

(ii) ***Technology issues*** decide whether current technology infrastructure can accomplish training goals, whether it justifies additional tech expenditure and whether software/hardware compatibility is possible.

1.4.3 Disadvantages to the Learner

(i) ***Technology issues*** of learners are usually technophobia/unavailability of needed technologies.

(ii) ***Portability*** of training is e-learning's strength due to proliferation of network linking points, notebook computers, PDAs and mobile phones.

Web-based learning environments (Kybartaite et al 2009) are of 2 types: synchronous and asynchronous. A synchronous learning environment is where an instructor teaches a traditional class with the instructor and students being online simultaneously, communicating with each other. Software tools for this learning type include audio conferencing, video conferencing, and virtual whiteboards ensuring that instructors and students share knowledge.

In asynchronous learning environment, instructor interacts with students intermittently and not in real time. Asynchronous learning is support technologies like online discussion groups, email, and online courses.

e-learning environments provide the following management, development and delivery of e-learning (Kybartaite, et al., 2009) capabilities:

(i) Map Competencies to Courses: An administrator knows competencies (skills) required for specific jobs in an organization; describing learning content (courses) that teach that skill.

(ii) Schedule Classes/Register Students: An administrator schedules synchronous classes/ posts links to asynchronous class courses. Students can register for either synchronous or asynchronous classes.

(iii) Track Learning: The system tracks classes a student takes and how he scores in class assessments.

(iv) Develop Learning Content: Authors are given software tools to create asynchronous courses consisting of reusable learning objects.

(v) Deliver Learning Content: Asynchronous courses or individual learning objects stored in the server are delivered to students via a Web browser client.

Collaborative issues in which e-learning communities unfold are characterized as complex as it requires negotiation/communication to uncover. It requires high reflexivity and involves collaborative (self/peer/tutor) assessment processes. Designing and facilitation of education approaches are sensitive to specific pedagogical cultures and educational traditions. Designing teaching and learning mediated across virtual/physical spaces in higher education contexts also come within its ambit.

1.5 INFORMATION RETRIEVAL (IR)

Information Retrieval (IR) includes locating unstructured (text) material (documents) which satisfies information need from large collections (stored on computers). Information Retrieval systems are distinguished by their operational scale. It is being useful to understand 3 prominent scales. The system searches more than a billion documents in millions of computers in a web search. Personal information retrieval is at the other extreme. Recently, consumer operating systems have integrated information retrieval.

The space for enterprise, institutional and domain-specific search (Dinh and Tamine 2012), where retrieval is possible for collections like a

corporation's internal documents, patents database or research articles on biochemistry is in between.

Information Retrieval is locating from a large collection, documents that fulfill a specific information need. Much Information Retrieval research concerns proposing and testing methodologies to perform this job. It can be considered that a formal relationships model between queries, documents, meaning and relevance can be used as a base for information retrieval. There can be no such model as humans cannot be left out of the equation, and yet cannot be modeled. Information retrieval techniques account for language, culture and behavior. For example, similarity estimation was circumscribed or bounded, as in cosine measure. Experiments are reliant on a standard test,but the aim is to measure users, and hence data over which prior knowledge could be asserted should be chosen so that it is not constrained by prior failures. This is important, as variables should not be present in systems: as effectiveness measurements must be controlled.

	Search by Navigation (following links as in a subject directory and the web generally)	Search by query (as in Google)	Creating answers and new information by analysis and inference-based on query
Unstructured information (text, images, sound)	**Hypermedia systems**(Many small units such as paragraph and single images tied together by links)	IR Systems (Often dealing with whole documents such as books and journal articles)	
Structured information		**Database Management Systems(DBMS)**	**Data analysis systems Expert Systems**

Figure 1.1The IR System Family

Two distinctions are of importance:

(i) An unstructured information system deals with issues like the Reformation's economic impact.

(ii) Finding versus creating answers. IR/database systems only locate what is already in existence.

IR has traditionally concentrated on locating entire documents which include written text; much IR research specifically focuses on text retrieval. IR is computerized retrieval of machine readable text sans human indexing.

Information retrieval systems deal with huge data amounts and should be able to process gigabytes or even terabytes of text. It should build and maintain an index for millions of documents.

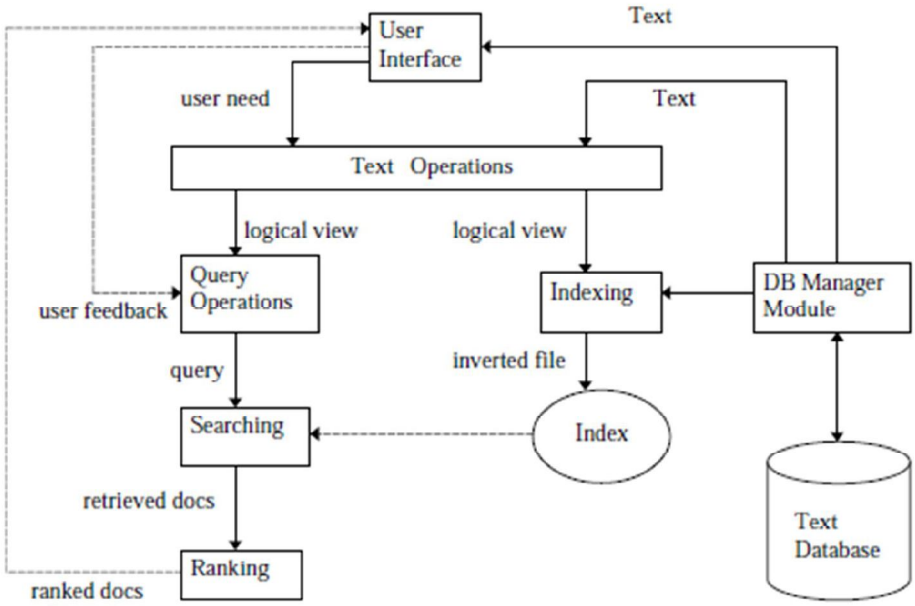

Figure 1.2 The process of retrieving information

1.6 INFERENCE NETWORK MODEL

Document retrieval is modeled through an inference process in an inference network, in this model. Most IR system techniques are implemented through this model (Boughanem, et al., 2009). In this model's implementation, a term is instantiated with certain strength by a document and credit from multiple terms is gathered following a query to compute a numeric score equivalent for the document. From an operational perspective, a term's instantiation strength for a document is considered the term's weight in the document, and simple document ranking in this model is similar to vector space model ranking and probabilistic models described. A term's instantiation strength for a document is not model defined and so any formulation can be utilized.

Natural Language Processing (NLP) enhances retrieval effectiveness, with limited success. Document ranking is a critical IR application, but it is not the only one as many techniques have been developed to attack varied issues including information filtering, topic detection and tracking (or TDT), speech retrieval, cross-language retrieval, question answering etc.

IR documents (Greengrass 2000) are partly structured, e.g., it has a structured header and unstructured body. The header has metadata, i.e., data about document, instead of the document's information content. For example, a book's structure consists of certain components due to it being a book, e.g., it contains title page, chapters, etc. IR retrieves documents based on unstructured components content. An IR request (called a "query") may specify a document's structured and unstructured components characteristics for retrieval.

1.7 INFORMATION RETRIEVAL SYSTEMS

IR tries to locate documents in a collection “about” a given topic or which satisfies a specific information need. Topic or information need is expressed through a user generated query. According to users, documents which satisfy a query are “relevant” and those not about a topic is “non-relevant”. A query may be used by an IR engine to classify a documents collection (or in an incoming stream), returning a documents subset which meets some classification criterion for the user. The higher the proportion of user returned documents that are relevant, the better the classification criterion (Kelly 2009). Figure 1.3 reveals a typical IR system.

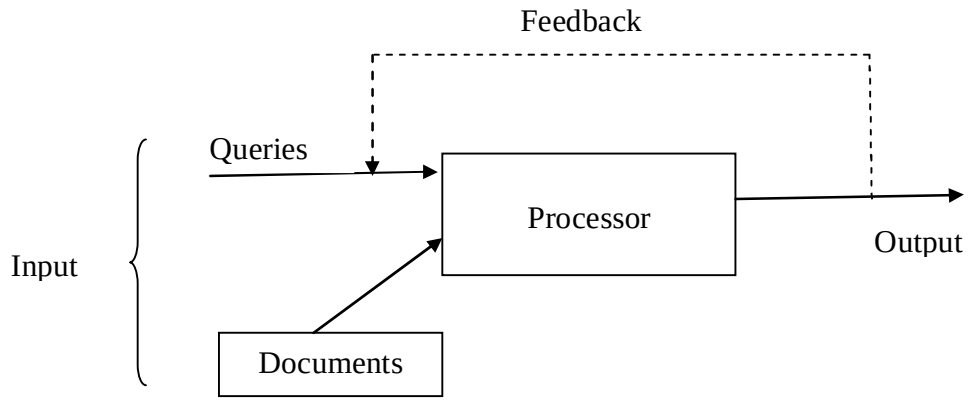

Figure 1.3 A Typical IR system

Relevance judgments effectiveness is quantified by two measures: recall (relevant documents number identified by a subject for a query divided by total relevant documents, within examined ones, for the query), and judgments precision.

Most automatic information retrieval systems are experimental in nature. Experimental IR is usually tried out in ‘laboratory’ situations whereas those which are operational are commercial, charging for their services. Both

systems are evaluated differently. 'Real world' IR systems are evaluated regarding 'user satisfaction' and the price a user pays for services. Experimental IR systems evaluation is by comparing retrieval experiments with specially constructed standards.

Retrieval Model

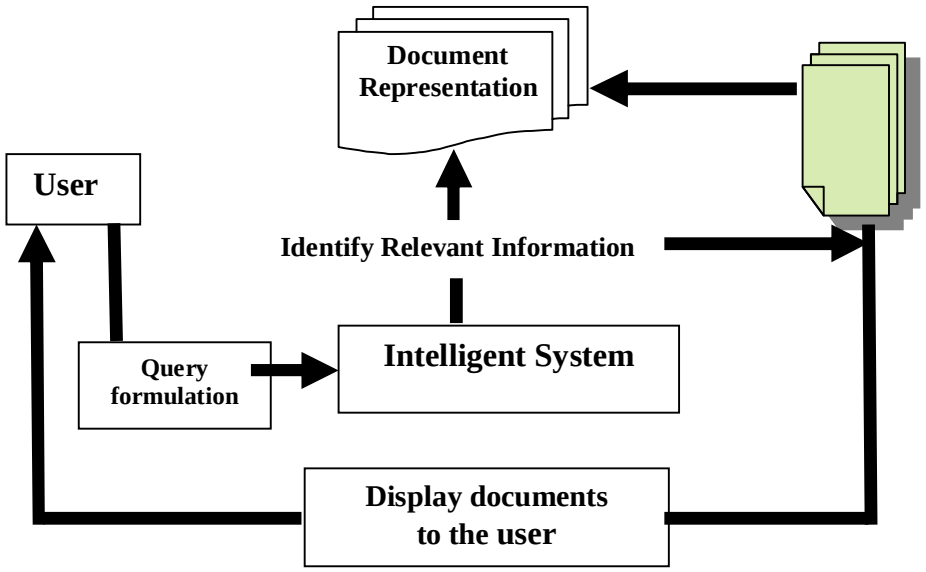

Figure 1.4 Retrieval models

Information storage and retrieval are simple in principle, if there is a document cluster and a person (user) formulates a question (request or query) for which a document set is the answer, satisfying his question's information need.

When high speed computers were available for non-numerical work, many felt a computer could 'read' an entire document collection to extract relevant documents. It soon was apparent that use of a document's natural language text not only resulted in input and storage problems, it also left unsolved document content characterizing problems. Future hardware

developments may ensure feasible natural language input and storage. Automatic characterization where software tries to duplicate human 'reading' process is a very sticky issue. Specifically, 'reading' involves extracting information - both syntactic and semantic - from text to decide whether a document is relevant to a particular request. The difficulty is in extracting information and using it to decide relevance. Modern linguistics slow progress on the semantic front and machine translation's failure reveal that such issues are yet unsolved.

Literature presents three IR models (Vinciarelli 2005): the first being called Boolean, the second is Vector Space Model (VSM) and the third is called probabilistic. Boolean model is based on binary algebra: queries being expressed as logical conditions. Probabilistic approaches estimate a document's probability being relevant to a specific query. This needs many training queries which are hard to get.

Documents are represented as vectors in Vector Space Model (VSM) and their relevance to user queries is measured by appropriate matching functions (Tsatsaronis and Panagiotopoulou 2009). There are two major components in the IR process: the first is term extraction by a document matrix performed for a given database, once. The second is document identification as relevant to a query and performed every time a query is submitted.

The term by document matrix 'A' is obtained through many steps: preprocessing, normalization and indexing. Preprocessing removes elements not useful in retrieval.

Normalization removes variability not useful to retrieval and is performed through two steps: stopping and stemming. During stopping, all words which have poor index terms (stopwords) are removed. Stemming is replacing different inflected forms of certain words with their stems.

In Information retrieval systems (Sanderson and Croft 2012) started electro-mechanical searching devices, adoption of computers to locate user query relevant items. Both electro-mechanical and computer-based IR systems search style is Boolean retrieval. A query is a logical term (a synonym to word in IR literature), which result in a documents set that match the query exactly. An alternative approach is where each collection of document is given, a score to indicate its relevance to a query. This ranked retrieval search approach was taken up by IR researchers, who over decades refined and revised documents sorting regarding a query. This approach effectiveness over Boolean search was proved over the years.

Information retrieval systems (Sanderson and Zobel 2005) effectiveness is measured by comparing performance on a common queries set and documents. Many tests evaluated such comparison's reliability.

Test collections are used for retrieval systems comparison and evaluation. These collections comprising of documents, queries (or topics), and relevance judgments are key Information Retrieval (IR) research for years; collections are based on research or practice in collection formation and retrieval effectiveness measurement. Effectiveness computation is madeby measuring systems relevant documents location ability. The measured score indicates a system's performance relative to another; it is being assumed that similar relative performance is observed on other test collections operational settings.

1.8 NATURAL LANGUAGE PROCESSING (NLP)

Natural Language Processing (NLP) (Krallinger et al 2005) is where computers are used to process language including techniques to provide basic methodology for automatic relevant information extraction from unstructured data, like scientific publications. Information retrieval and NLP systems are likely to become important for both information extraction and assisting in research's various aspects like new facts discovery, findings interpretation, and experiments design. Though useful for many jobs, such tools consume time when used for efficient searches and article selection, such functions being repeated regularly to update knowledge.

Identification of entities in free text in NLP is called Named-Entity Recognition (NER). To identify biological entities like genes, proteins and drugs automatically within free text, over fifty information-extraction/text-mining tools were recently implemented with two community-wide evaluations being carried out.

When a document collection is obtained, clues are given for documents to be retrieved from a collection. Then documents matching clues are answers to a query. In invoking a search engine, a few words are presented which are matched to stored documents, the best matches being the responses. The process is generalized to a document matcher, where instead of few words a complete document becomes the clue. Input document is matched with all stored documents, with best matching documents being retrieved.

A basic information retrieval concept is measuring similarity between two documents. For this, a small word set input into a search engine

is considered, a document to be matched to others. Measuring similarity is related to predictive methods for learning/classifying methods. Measuring similarity is a common theme, and the method's variations are basic to information retrieval.

1.9 MOTIVATION AND PROBLEM STATEMENT

The advent of internet technology to share educational contents and experiences ensured that institutions globally, offered a federated search to courses, lesson plans, contents, assignments, seminars and experiments, all of which are stored in repositories of learning content management systems controlled by a learning management system. The problem faced by a learner's community is in accessing, sharing and delivering quality relevant to content for online teaching learning systems. Today peer to peer networks are used for daily sharing of videos, audios, images, music or other distributed learning digital processes. Hence, sophisticated search and information retrieval solutions are necessary. Many existing web based course structures developed in e-learning system are considered as course ontology and can be mapped into a model. This ontology based solution increases information retrieval accuracy through high precision and recall.

Language Model defines documents probability distribution using them to predict likelihood of query terms observation. Language model has been defined for all documents and it is used to inquire about chances of query generation. Nominal Language Model (NLM) based language modelling goes with part of speech of a given query's literal language, constituting factors with noun and adjectives. Informational query attempts to capture a document with data, relevant to analysis area. NLM based Information Retrieval process is an efficient method to extract relevant

documents. Language modelling is processed with natural language processing methods.

A term/phrase can have many meanings, while a domain specific concept is unambiguous. It is useful to use the domain specific concepts in documents than terms to retrieve documents from a specific domain. In this proposal, NLM is assembled with rate specifications and ratio calculations through use of probabilistic terms involving comparing query terms occurrence with data store using conditional probability theorem.

Feature selection in classification is being viewed as a most fundamental issue in machine learning. Data clustering is a popular data labeling technique where unlabeled data is issued, and similar samples put in one pile called a cluster with dissimilar samples being in other clusters. Data clustering, an NP-complete problem of locating groups in heterogeneous data by minimizing some dissimilarity measure is a basic tool in data mining, machine learning and pattern classification solutions.

Query expansion methods were long studied - with debatable success on many occasions. This study presents a probabilistic query expansion model based on similarity thesaurus constructed automatically. The latter reflects domain knowledge about specific collections from which it is constructed. Two important issues with query expansion are addressed here: selection and weighting of additional search terms. Compared to earlier methods, queries are expanded by addition of terms similar to query concept, instead of selecting terms similar to query terms.

Mobile Agents are independent smart programs moving through networks, seeking and interacting with available/compatible services on user’s behalf. Another attractive paradigm property is that it allows an application to be really distributed, as tasks in an application, embodied in a

mobile agent, are worked out on participating systems in a decentralized process.

This study addresses pre-processing and various source documents retrieval to achieve improved Information Retrieval systems and investigates tools/techniques used for autonomous classification or documents clustering; new methods are proposed based on concept expansion.

1.10 OBJECTIVES OF THE THESIS

Locating learning material groups relevant to learning goal (query), results in learning process efficiency. The scope of this thesis is in increasing content retrieval efficiency and accuracy through a query refining and reformulation method through pre-processing operations like stemming, stop word removal, dimensionality reduction and relevance feedback mechanism. Reuter's dataset and Movielens dataset are used in this research. The following summarizes the objective of the thesis:

(i) Propose a concept expansion for creating corpus

(ii) Propose a language modeling based on Nominal Language Model

(iii) Propose a cluster based Feature selection method based on Particle Swarm Optimization and Genetic Algorithm

(iv) Optimization of document clustering using proposed Mobile Agent Figure 1.5 shows the flow of the proposed methodology.

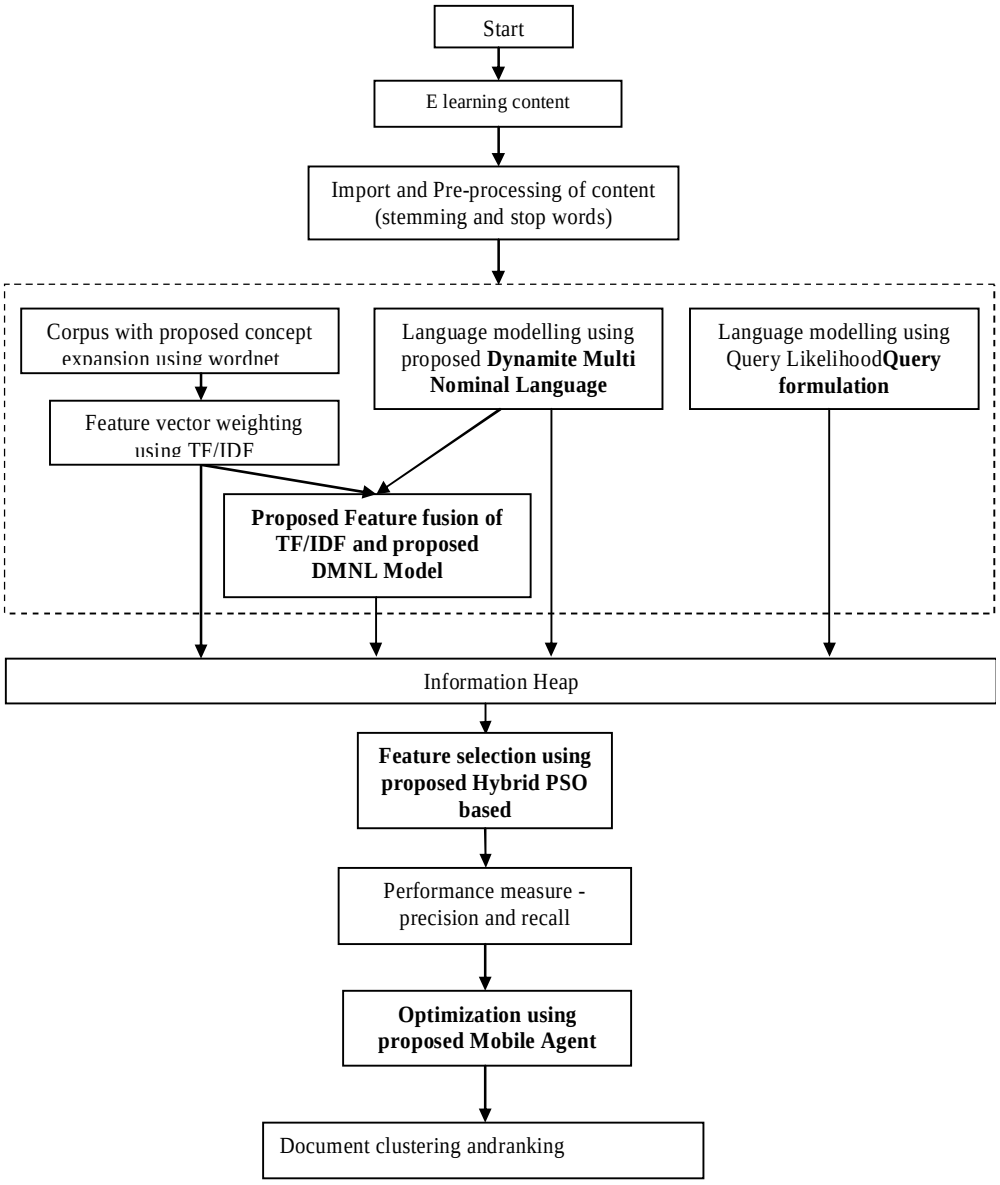

Figure 1.5 Flowchart of the Proposed Methodology

1.11 ORGANIZATION OF THE CHAPTERS

The second chapter deals with literature survey of work related to e-learning, document clustering and Ranking, IR Systems, concept based approaches in IR and language models.

The third chapter discusses in detail about various language modeling. A concept expansion for creating corpus is proposed. The evaluation on the proposed method using Reuter's dataset and MovieLens data is discussed.

The fourth chapter details with the proposed language modeling based on nominal language model.

The fifth chapter proposes a cluster based feature selection method based on Particle Swarm Optimization (PSO) and Genetic Algorithm (GA).

The sixth chapter discusses the process of optimization of document clustering using proposed mobile agent. The experimental results are explained in a detailed manner.

The seventh chapter concludes the research work.