CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

This chapter deals with the review of various aspects of e-learning and Information retrieval. The papers reviewed are sectioned as follows and tabular version has been enclosed in the Annexure I and II:

(i) e-learning
(ii) Document clustering and ranking
(iii) Information retrieval systems
(iv) Concept-based approaches in information retrieval
(v) Language models

2.2 e-LEARNING

Strijbos et al (2011) discussed addressing collaborative learning assessment with a perspective on what is to be assessed, highlighting current approaches and limitations. Within Computer-Supported (CS) Collaborative Learning (CL) research community; there was major dialogue on theories/perspectives on collaborative learning, scripting the collaborative process approaches, with research methodology being the most recent. Assessment of collaborative learning received less attention in contrast. As collaborative learning assessment is demanding for teachers and students adequate
computer-supported and intelligent tools to monitor and assess are needed. A roadmap is presented as regards the role and application of intelligent tools for assessment of CL.

Brut et al (2011) extended IEEE LOM standard with ontology-based semantic annotations solution for use of learning objects sans Learning Management Systems. The data model which corresponds to this approach is presented first. Also, presented is a proposed indexing technique for this model’s development to acquire a better learning resources annotation. Two alternative methods for structure-based indexing of textual resources are extended and combined: the latent semantic indexing mathematical approach and linguistic-oriented WordNet-based text processing. Hence, the reason behind good results due to the first method becomes transparent, because of linguistic controlled choices suggested by the second method. In the context of adopting semantic web technologies for e-learning, the results are important, and also as progress in ontology-based textual resources indexing.

An agent-based m-learning system was presented by Glavinic et al (2008). This study provides a roadmap for m-learning systems design based on agent technology. M-learning, as a "portable and personal" e-learning fashion, enhances the efficiency/effectiveness of learning in hand held terminals. Mobile intelligent tutoring systems are specific m-learning systems basing their work on a human teacher simulation in a learning/teaching process. Systems should provide intelligent support to learners and adapt to various networks/mobile devices. Consequently this implies use of agent-based solutions as an architectural basis. The issue is lack of agent-based software development methods and common scenarios/templates to build multi-agent systems.

SWAM, a platform to build and deploy Prolog-based intelligent mobile agents on Semantic Web was presented by Crasso et al (2011). A big
repository of static data, the Web has gradually become a worldwide network of information and services called Semantic Web which enables programs to interact autonomously with Web-accessible information/services. Thus, mobile agent technology helps efficiently exploit the new Web in an automated way, as Semantic Web resources are described in a computer-literate way. This article reports examples and experimental results to illustrate and assess SWAM’s benefits.

Cuéllar et al (2011) developed a procedure to integrate different e-learning systems, and give semantics to Learning Management Systems (LMS) database entities or relations through ontology. Internet LMSs are tools that help in daily teaching/learning activities. Most users or the software is focused on content dissemination and group works, but internet LMS possibilities could go further. Recent approaches use semantic web to improve e-learning capabilities and user experiences through artificial intelligence and knowledge management techniques. The suggested integration eases learning resources and knowledge dissemination from LMS databases. Also, database schemes’ semantic interpretations ensure location of precise information expeditiously.

A performance-oriented e-learning approach was presented by Jia et al (2011). Despite increasing workplace e-learning practices, most e-learning applications do not meet learners’ needs or serve organization success quest. Major gaps between organizational interests and individual needs exist in e-learning, making such applications less goal-effective. To overcome this, Key Performance Indicators (KPIs) clarify organizational training needs in addition to helping learners establish rational learning objectives. Ontology is also used to construct formal and machine-understandable performance-oriented learning environment conceptualization.
With this approach, a KPI-oriented learning ontology and prototype system were developed /evaluated to demonstrate the approach’s effectiveness.

2.2.1 Evaluation of e-learners Standards

Like all technology e-learning has evolved standards from 1994 to develop a metadata driven framework to access web resources due to discovery of electronic resources in 1997.

The National Institute for Standards and Technology (NIST) and the IEEE p.1484 study group (IEEE-LTSC-IEEE learning technology stands technology) collaborated with ARIDIVE a European project having active metadata definitions.

In 1998, Instructional Management System(IMS) and ARIADIVE submitted joint proposal/specifications to the IEEE Learning Object Metadata (LOM) base document.

2.2.2 Aviation Industry CBT Committee (AICC)

AICC is an information training professional forum to develop guidelines in the aviation industry for the development, delivery and evolution of CBT and related training technology.

AICC objectives are as follows

(i) Developing guidelines for interoperability

(ii) Providing open forum to describe CBT and training technology.

(iii) Promoting economic/effective computer based training.

(iv) Focusing on online learning reuse/interoperability.
(v) Coordinating effort with learning technology standards like Learning Technology Systems Architecture (LTSA) of IEEE and Advance Distributed Learning (ADL).

2.2.3 Dublin Core Metadata Initiatives (DCMI)

DCMI is an organization promoting adoption of interoperable metadata standards and developing specialized metadata vocabulary to describe resources enabling more intelligent information discovery systems. The fifteen elements of Dublin core semantics were established by an international, cross disciplinary group of professionals through consensus, from librarianships, computer science, text encoding, museum community and related fields.

The fifteen elements are as follows

(i) Title: A Title is a name by which resources are formally known.

(ii) Creator: An entity responsible to make resource content. For example, authors in written documents, artists, photographers, or illustrators in visual resources.

(iii) Subject and Keywords: The content topic of resource. A Subject is expressed as keywords, key phrases or classification codes describing a resource topic.

(iv) Description: An account of resource content. Description includes but not limited to, abstract, contents tables, reference to graphical content representation or content’s free-text account.
(v) Publisher: An entity responsible to make resource available.

(vi) Contributor: An entity to make contributions to resource content.

(vii) Date: Date will be associated with resource creation/availability. Recommended best practice to encode date value is defined in ISO 8601 profile and follows YYYY-MM-DD format.

(viii) Type: Nature/genre of resource content. Type includes content terms describing general categories/ functions/genres/aggregation levels.

(ix) Format: This includes media-type/resource dimensions. Format may determine software/hardware or other equipment to display/operate resource. Dimension examples include size and duration.

(x) Identifier: A reference to resource in a given context. Recommended best practice is identifying resource through a string/number in a formal identification system. Examples include Uniform Resource Identifier (URI) including Uniform Resource Locator (URL), Digital Object Identifier (DOI) and International Standard Book Number (ISBN).

(xi) Source: A reference to resource from which present resource is derived.

(xii) Language: Language of the intellectual resource content. Recommended best practice is using RFC 3066 in conjunction with ISO 639 to define two- and three-letter primary language tags with optional subtags. Examples
include "en" or "eng" for English, "akk" for Akkadian, and "en-GB" for English used in the UK.


(xiv) Coverage: Extent/scope of resource content. Coverage includes spatial location (place name/geographic coordinates), temporal period (period label, date/date range) or jurisdiction (named administrative entity).

(xv) Rights: Rights Information held in and over resource. A Rights element contains a resource’s rights management statement or refers to a service giving such information. It also encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights.

2.2.4 ARIADNE:

ARIADNE is a European digital library project for academic and corporate context providing access to resources and focused on tools and methodologies development to produce, manage and reuse computer based pedagogical elements based on syllabi and curricular. Ariadne and IMS jointly developed a metadata specification for submission to IEEE.

2.2.5 Advance Distributed Learning initiative (ADL)

This is US department of defense and white house office of science and technology program to develop guidelines for large scale development/implementation of efficient/effective distributed learning.
2.2.6 Sharable Content Object Reference Model (SCORM)

The US Federal government ADL initiative and recently relearned SCORM are recent examples of learning standards application and integration. This provides a foundation for learning techniques to build Operate in a future learning environment. Military, Air, Navy, Army or Marine can exchange, manage, track and review all learning content/date whatever be its source/application.

2.2.7 Instructional Management System (IMS) Global Learning Consortium

In 1997, the IMS project focused on initiatives relating to standards for learning servers, learning content and enterprise integration of such capabilities.

IMS Consortium is at an advanced stage of developing learning resources metadata specification for creation of a uniform way to describe learning resources to ensure easy focus.

To share learner’s data, courses perform across platforms, operating system and user interfaces. Content Packaging Specification ensures easier creation of reusable, sharable content objects used across many learning management systems. Question and test specification address the need to share test items, and assessment tools across varied systems. IMS learner profile specification organizes learner information to ensure that learning system is responsive to learner’s specific needs.

2.2.8 IEEE Learning Technology Station Committee (LTSC)

The IEEE Computer Society Standards Activity Board chartered LTSC develops accredited technical standards, recommended practices/guides
for learning technology. LTSC formally/informally coordinates with other organizations which produce specifications/standards for similar purposes. Up to twenty working groups discussed various e-learning aspects at various times.

IEEE LTSC has over a dozen working groups and study group.

(i) IEEE 1484.1 Architect and ref model
(ii) IEEE 1484.3 Glossary
(iii) IEEE 1484.2 Learner model
(iv) IEEE 1484.13 Student identifier
(v) IEEE 1484.19 Quality system for lifelong learning
(vi) IEEE 1484.10 CBT Data Exchange
(vii) IEEE 1484.6 Course Sequencing
(viii) IEEE 1484.17 Content Packages
(ix) IEEE 1484.12 Learn Object metadata
(x) IEEE 1484.9 Localization
(xi) IEEE 1484.14 Semantics and Exchange bindings
(xii) IEEE 1484.15 Date Interchange Protocols
(xiii) IEEE 1484.11 Computer managed lasting.
(xiv) IEEE 1484.18 Tool/Agent Communications
2.2.9 Learning Object Metadata (LOM)

LOM is an approved IEEE-SA standard passed on to ISO/IEC JTC1/SC36 to convert to international standards. LOM is based on ARIADNE, IMS and DCMI defining a structure for different granularities learning objects interoperable descriptions. A learning object is an entity digital/non-digital that is used to learn, educate or train. LOM descriptions are grouped as general, life cycle, meta-metadata, educational, technical, educational rights, relation, annotation, and classification. It doesn’t define how learning technology systems represent/use metadata instance to learn an object; as partly defined in IMS and ADL/SCORM.

This standard’s purpose is facilitating learning objects search, evaluation, acquisition, use, sharing and exchange by learners/ instructors/automated software processes like course authoring/structuring tools.

LOM defines nine categories to group various data elements (Hodgins and Wason 1998):

(i) General category groups: general information describing learning object as whole, like title, language, description, keyword, coverage, structure (underlying learning object’s organizational structure; e.g. atomic, linear or hierarchical) and aggregation level (granularity level – from raw media to a course set)

(ii) Lifecycle category groups: features related to learning object’s history and current state and those affected by learning object during evolution. Elements include version, status, and state of object contributors.
(iii) The Meta-Metadata category groups: metadata instance information (rather than learning object described by metadata instance). E.g. a unique record identifier, metadata contributors, metadata schema (e.g. LOMv1.0) and language.

(iv) The Technical category groups: Learning object’s technical requirements/characteristics. Elements are format (mime type), size, location (URL or URI), technical requirements, installation remarks, other platform requirements and duration.

(v) The Educational category groups: Learning object’s key educational and pedagogic characteristics. This includes interactivity type (active learning, like an exercise/simulation vs. expositive=passive learning, like reading), learning resource type (exercise, simulation, questionnaire, diagram, figure, graph, index, narrative text), interactivity level (very low – very high), semantic density (conciseness degree), intended end user role (teacher, author, learner, manager), context (school, higher education, training), typical age range, typical learning time, description and language.

(vi) The Rights category groups: Learning object’s intellectual property rights and usage conditions. It includes a cost field (yes/no), copyright (yes/no), and description.

(vii) The Relation category groups: Learning object’s features defining a relationship with other related learning objects like kind (nature of relationship, e.g. is/has part of, is/has version of, is/has format of, is referenced by, is based on, is basis for, requires, is required by), resource (target learning object resource) and description.
Annotation category gives comments on learning object’s educational use providing information on when and by whom comments were created. It has an entity (people, organisation who created annotation), date and description.

Classification category describes learning object with regard to specific classification system. Elements are purpose (skill level, competency, security level, educational level, discipline, idea, prerequisite, educational objective, accessibility and restrictions), a specific classification system’s taxonomic path, description and keywords.

This standard facilitates search, evaluation, acquisition, use, sharing and exchange of learning objects by learners/instructors or automated software processes like course authoring and structuring.

2.2.10 Learning Technology System Architecture (LTSA) – Reference Model

This specifies a high architecture level for technology enhanced learning, education and training systems to

(i) Provide a framework to understand existing/future systems

(ii) Promote interoperability and portability through identifying abstract, high level system interface.

(iii) Incorporate a technical horizon of 5-10 years minimum while being adapting new technologies/learning technology system.

Standardisation document clarifies five refinement architecture layers, but layer three alone is normative, the remaining four layers being meant for information and completeness.
The system’s five layers as seen in Figure 2.1 separate the “big picture” from “details” and help understand/analyze the system step by step. Each layer is investigated independently as they do not influence each other. They are called:

(i) **Learner and Environment Interactions**: Concerns learner's acquisition, transfer, exchange, formulation, discovery, etc., of knowledge and/or information through interaction with environment from information technology perspective and not a description of learning theory. It shows the learner has new/different knowledge after a learning experience.

(ii) **Learner-Related Design Features**: is about the effect learners have on learning technology systems, and how it is affected by learners needs specifically, the nature of human learning.

(iii) **System Components (normative)**: Describes component-based architecture, as identified in human-centred and pervasive features. The LTSA identifies
a) four processes: learner entity, evaluation, coach, and delivery process;

b) two stores: learner records and learning resources;

c) thirteen information flows among components: behavioural observations, assessment information, learner information (thrice), query, catalogue info, locator (twice), learning content, multimedia, interaction context, and learning preferences.

Figure 2.2 The LTSA system components

The operation has the form (Figure 2.2):

a) learning styles, strategies, methods are negotiated among learner and stakeholders and communicated as learning preferences;

b) learner is observed and evaluated regarding multimedia interactions;
c) evaluation produces assessments and/or learner information;

d) learner information is stored in learner history database;

e) coach reviews learner’s assessment/information like preferences, past performance history, and, possibly, future learning objectives;

f) coach searches learning resources, through query/catalogue info, for learning content;

g) coach extracts locators from catalogue info passing locators to delivery processes, e.g., a lesson plan; and

h) Delivery process extracts learning content from learning resources and based on locators transforms learning content to learner through an interactive multimedia presentation.

(iv) Implementation Perspectives and Priorities: Describes learning technology systems from various perspectives through reference to system components layer subsets. Different use of e-learning systems case models are analyzed and inter-process and communication models sketched.

(v) Operational Components and Interoperability — coding’s, APIs, protocols: Describes generic "plug-n-play" (interoperable) components/interfaces of information technology-based learning technology architecture, as seen in stakeholder perspectives.
Actual coding specification and API protocols standards are outside LTSA scope.

### 2.2.11 Platform and Media Profiles

This work group aims to identify other standards/ formats relevant to e-learning and browser platforms/media types. Standard does not describe technical details, only limitations/enhancements to referring standards. This standard specifically deals with the following issues:

(i) 1484.18.1.*: Bundles of profiles: e.g. a browser with specific capabilities set (JavaScript, Java, HTML, CSS and media type support) and plug ins

(ii) 1484.18.2.*: Markup Languages: various HTML, XML and style sheet versions.

(iii) 1484.18.3.*: Audio Formats: like wav, real audio and mp3.

(iv) 1484.18.4.*: Video and Graphics Formats: e.g. avi, quicktime, mpeg, jpeg, gif, bmp, png, flash, shockwave, cgm.

(v) 1484.18.5.*: Page Description Languages: e.g. PDF and Postscript.

(vi) 1484.18.6.*: Java: various JDK and JVM versions.

(vii) 1484.18.7.*: JavaScript: various JavaScript and ECMA script versions.

(viii) 1484.18.8.*: Word Processing Formats: e.g. RTF, Microsoft Word, WordPerfect etc.
(ix) 1484.18.9.*: Presentation Graphics: e.g. Microsoft PowerPoint.

(x) 1484.18.10.*: Spread sheet Formats: e.g. Microsoft Excel.

(xi) 1484.18.11.*: Document Services currently refers to DOM level 1.

2.3 DOCUMENT CLUSTERING AND RANKING

Sanz-Rodríguez et al (2010) tried improving recommending learning objects, this document highlighting insufficiencies of current approaches, identifying quality indicators to provide information based on which materials can be recommended to users. Next, a synthesized quality indicator to facilitate learning objects ranking, according to overall quality, is suggested. Thus, clear evaluations by users/experts are used with usage data thereby completing recommendation based information. The relationships that exist between different quality indicators in a set of learning objects from the Merlot repository were analyzed to ensure an overall quality indicator to be calculated automatically, guaranteeing rating of all resources.

Ochoa et al (2008) improved learning object search’s current status. First, the situation is analyzed to propose a theoretical solution, based on relevance ranking. This study develops relevance concept in a learning object search context to implement the solution. Based on the concept, a set of metrics to estimate topical, personal and situational relevance dimensions is proposed. Metrics are calculated from usage and contextual information not needing any user information. An exploratory metrics evaluation reveals that even simple ones provide statistically significant ranking order improvement, over a most common algorithmic relevance metric. Also, combining metrics
through learning algorithms sorts result list better by 50 per cent than baseline ranking.

Hristidis et al (2011) presented algorithms to return query based top results, ranked according to an IR-style ranking function, when operating on a Boolean query source interface with no ranking capabilities (or a ranking capability that did not interest end user). Many online/local data sources provide good querying mechanisms with reduced ranking facilities. For instance, PubMed permits users to submit expressive Boolean keyword queries, but ranks query results only according to date, but users prefer relevance ranking, measured by an information retrieval (IR) ranking function. A naive approach would be to submitting a disjunctive query with query keywords and to retrieve all returned matching documents, and then re-rank them. This would be expensive due to large number of disjunctive query returned results. The proposed algorithms enable conjunctive queries which return only candidate documents for relevance metric based high ranking. This process can also be applied to settings where ranking is monotonic on a factors set (query keywords in IR) and source query interface is their Boolean expression. A comprehensive evaluation on PubMed database and a Text REtrieval Conference (TREC) data set prove achievement of high order magnitude improvement when compared to existing baseline approaches.

Yen et al (2010) followed SCORM and Content Object Repository Discovery and Registration Architecture (CORDRA) specifications in developing a registry system, MINE Registry. Based on internet and search engine popularity, users request information through web based services. Though general-searching as provided by Google is powerful, metadata is needed as a searching mechanism for specific purposes. SCORM ensures efficient metadata definition for learning objects to be searched and shared in e-learning. To enable a federated repository search, CORDRA ensures a
common architecture to discover/share Learning Objects. MINE Registry stores and shares 20,738 Learning Objects generated over the last five years. The concept of “Reusability Tree” to represent relevant Learning Objects relationships was proposed and enhanced CORDRA as a contribution. Relevant information was collected, when users utilized Learning Objects like citations and time period persisted. Feedbacks from users are critical elements to evaluate significance degree of Learning Objects. A mechanism to weight/rank Learning Objects in MINE Registry was proposed through such factors, in addition to external learning objects repositories. A tool called “Search Guider” is provided to assist users find relevant information in individual requirements based Learning Objects.

A new spectral clustering method called Correlation Preserving Indexing (CPI) was presented by Zhang et al (2012). This is performed in correlation similarity measure space where documents are projected into a low-dimensional semantic space within which correlations between documents in local patches are maximized while correlations between documents outside the patches are reduced simultaneously. As document space’s intrinsic geometrical structure is embedded in similarities between documents, correlation as a similarity measure suits detection of intrinsic geometrical structure of document space more than Euclidean distance. Hence, the proposed CPI method effectively discovers intrinsic structures fixed in high-dimensional document space. The new method’s effectiveness is demonstrated by experiments conducted on various data sets and comparing them with current document clustering methods.

Duh et al (2011) examined whether easy to obtain additional unlabelled data can improve supervised algorithms. Ranking functions are important in information retrieval systems. Recently there was much research in “learning to rank,” which aims to use labelled training data and machine
learning algorithms to ensure construction of reliable ranking functions. Machine learning methods like neural networks, support vector machines, and least squares were successfully applied to ranking problems, and some were already deployed in commercial search engines. Despite this, most algorithms today construct ranking functions in a supervised learning setting, assuming that relevance labels are ensured by human annotators before training a ranking function. Such methods perform poorly when human relevance judgments are unavailable for a query range. The transductive setting was specifically investigated, where unlabelled data is equal to test data. A simple, flexible transductive meta-algorithm was proposed, the idea being to adapt training procedure to test lists after observing documents that need ranking. Two instantiations of this general framework were investigated: Feature Generation approach discovers more salient features from unlabelled test data and trains a ranker on a test-dependent feature-set. The important weighting approach relies on ideas in domain adaptation literature, and works by training data re-weighting to match statistics of each test list. Both proposed approaches proved that it performs better than supervised algorithms on TREC and OHSUMED tasks from a LEarning TO Rank (LETOR) dataset.

Cai and Li (2013) proposed an approach which directly generates clusters that were integrated with ranking. The idea was that ranking distribution of sentences in each cluster should be quite different from each other, which may serve as features of clusters and new clustering measures of sentences was calculated accordingly. Also better clustering results achieves better ranking results too. As a result, performance of ranking and clustering is improved.

Navaneethakumar(2013) proposed a conceptual rule mining on text clusters to evaluate the more related and influential sentences contributing the document topic. Author has plan to extend conceptual text clustering to web
documents, by assigning the sentence weights based on conditional probability. With sentence rank conceptual rules, the text cluster documents were defined. The conceptual rule depicts the finer tuned document topic and sentence meaning utilized for evaluating the global document contribution.

2.4 INFORMATION RETRIEVAL SYSTEMS

A framework to build an adaptive Learning Management System (LMS) was suggested by Yaghmaie et al. (2011) based on multi-agent systems and using both Sharable Content Object Reference Model (SCORM) 2004 and semantic Web ontology to ensure learning content storage, sequencing and adaptation. This system was implemented on an open-source LMS and its functionalities demonstrated through simulation of a scenario mimicking real life conditions. The result proves the system’s effectiveness and appears very promising.

Larkey et al. (2005) proposed two probabilistic approaches. Cross-lingual retrieval is used today and is based on probabilistic relevance models, as exemplified by INQUERY, and on those based on language modeling. INQUERY, a query net model, ensures easy incorporation of query operators, including a synonym operator, and proved to be very useful in Cross-Language Information Retrieval (CLIR); an approach termed structured query translation. In sharp contrast, language models include translation probabilities in a unified framework. With Arabic and Spanish data sets, two approaches were compared with two kinds of bilingual dictionaries—one from a conventional dictionary, and another from a parallel corpus. Structured query processing provided slightly better results when queries were not expanded, but when queries were expanded, language modeling ensures better results, but again only when using a probabilistic parallel corpus dictionary. Two additional issues in the structured query processing comparison with language modeling were pursued. The first was regarding query expansion
and the second, the role of translation probabilities. Conventional expansion techniques (pseudo-relevance feedback) was compared with relevance modeling, a new IR approach that fit into formal language modeling framework. Relevance modeling and pseudo-relevance feedback achieved comparable retrieval levels with good translation probabilities conferring a small but significant advantage.

A web-based learning support system that harnesses two approaches - learning path constructing approach and learning object recommending approach was developed by Hsieh et al (2010). Recently, browsers are popular tools for internet information searching. Though users can locate and download specific learning materials to gain fragmented knowledge, most materials are imperfect without specific content order. This results in self-directed learners spending more time surveying and choosing right learning materials from the Internet. The system discovers candidate courses using data mining based on Apriori algorithm founded on collected documents and a learning subject from learners. Next, the Formal Concept Analysis (FCA) based leaning path constructing approach, builds a Concept Lattice with keywords got from selected documents. Then a relationship hierarchy of all the concepts represented by keywords is formed. It then uses FCA to compute more mutual relationships among documents to ensure a correct learning path. The support system uses both preference-based and the correlation-based algorithms to recommend most suitable learning objects/documents for a course’s units to ensure efficient learning for learners. Such an e-learning support system is capable of being embedded in any information retrieval system for surfers to ensure better and efficient Internet learning.

Ko et al (2008) proposed a helpful snippet generation method which uses a statistical query expansion approach with pseudo-relevance feedback and text summarization techniques being applied to salient sentence
extraction for quality snippets. A (page/web) snippet is a document excerpt which allows a user to know if a document is relevant without accessing it. In experiments, the proposed method proved to have much better performance than other methods including commercial Web search engines like Google and Naver.

Na et al (2007) examined usage of parsimony in query expansion and clustering algorithms effect in cluster-based retrieval. The term mismatch problem in information retrieval is critical with many techniques like query expansion, cluster-based retrieval and dimensionality reduction being developed to solve it. An empirical study on query expansion through cluster-based retrieval was undertaken. Again, query expansion and cluster-based retrieval are compared and combinations evaluated regarding retrieval performance through experiments on seven NTCIR and TREC test collections.

An effective technique to improve retrieval effectiveness, Relevance Feedback (RF) was studied in monolingual and Trans-Lingual Information Retrieval (TLIR) by He et al (2010). RF studies in TLIR were focused on Query Expansion (QE), where queries were reformulated before/after translation. RF in TLIR not only selected better query terms, but also enhanced query translation through adjusting translation probabilities and resolving out-of-vocabulary terms. This study proposes a novel relevance feedback method, Translation Enhancement (TE), using extracted translation relationships from documents to revise query terms translation probabilities and to identify extra translation alternatives to ensure that translated queries are tuned to current search. TE was studied with Pseudo-Relevance Feedback (PRF) and Interactive Relevance Feedback (IRF) and the results revealed that TE significantly improved TLIR with both relevance feedback methods, and that improvement was comparable to query expansion. Also, translation
enhancement and query expansion effects were complementary, their integration further improving TLIR to be more robust for various queries.

Lin et al (2006) presented a new user relevance feedback based query expansion method to mine additional query terms. Proper query terms greatly affect document retrieval systems performance. System performance is improved by using query expansion techniques. According to user’s relevance feedback, the proposed method calculates relevant terms of documents degrees of importance in the document database. Relevant terms have higher importance degrees and may become additional query terms. The proposed method uses fuzzy rules to infer additional query terms weights. Then, additional query terms weights and original query terms weights form a new query vector used for document retrieval. The proposed method increases information retrieval systems’ precision and recall rates when handling document retrieval.

Mei et al (2007) proposed and studied a Poisson distribution based new family of query generation models. Many language model variants were proposed for information retrieval and most are based on multinomial distribution scoring documents on query likelihood computed on a query generation probabilistic model. It was shown that in their simplest forms, the new models family and existing multinomial models are both equal, but behave differently for smoothing methods. The Poisson model has many advantages over multinomial model. This includes accommodating per-term smoothing and ensuring more accurate background modeling. The new models variants were presented corresponding to different smoothing methods and were evaluated on four representative TREC test collections. Results revealed basic models to perform comparably, while the Poisson model outperformed multinomial model with per-term smoothing. Improved performance is possible with two-stage smoothing.
Kim et al (2011) presented novel algorithms to extract templates from many web documents generated from heterogeneous templates. World Wide Web is a useful information source. To achieve high publishing productivity, web pages in many websites are populated automatically by common content templates which provide readers access to contents through consistent structures. For machines, templates are harmful as it degrades web applications accuracy and performance due to the presence of irrelevant terms. Thus, template detection techniques gained attention recently to improve search engine performance, clustering, and web documents classification. Web documents were clustered on underlying template structures similarity in documents to ensure simultaneous extraction of templates for each cluster. A novel goodness measure was developed with approximation for clustering providing a comprehensive analysis of this algorithm. Experiments with real-life data sets confirm the algorithm’s effectiveness and robustness compared to state-of-the-art template detection algorithms.

Zidi and Abed (2013) presented a generic framework for ontology-based information retrieval. Recognition of semantic information extracted from data sources and the mapping of this knowledge into ontology has been focused. Semantic indexing based on entity retrieval model has been proposed to achieve high scalability. Ontology of public transportation domain was used in order to validate these proposals. Finally, this system uses ontology mapping and real world data sources.

Xue and Yan (2012) made a research on multi agents information retrieval system based on the intelligent evolution which includes nine different modules such as user Agent, communication Agent, mining Agent, personal Agent, intelligence evolution Agent, information retrieval Agent, group Agent, and intelligence evolution filtering Agent and clustering Agent.
Based on the user history query information, current retrieval information and feedback information rectifies the judgment of the user preferences constantly, makes the returned query results.

2.5 LANGUAGE MODELLING

Language modeling approach to information retrieval was promising as it linked retrieval with language model estimation, studied extensively in applications like speech recognition. The basic idea was to estimate a language model for a document, and rank it according to query likelihood of estimated language model. Language model estimation’s main problem was smoothing, which adjusts maximum likelihood estimator correct data sparseness related in accuracy. Language model smoothing and its retrieval performance influence were studied. Retrieval performance sensitivity to smoothing parameters was examined and compared to many smoothing methods on varied test collections.

Lafferty and Zhai (2001) presented an information retrieval framework combining document and query models through a probabilistic ranking function founded on Bayesian decision theory. The framework suggested operational retrieval extends language modeling approach developments to information retrieval. Each document’s language model was estimated, and also a language for each query with retrieval being cast in terms of risk minimization. Query language model is exploited for user preference modeling as regards query, synonymy and word senses contexts. Recent work incorporated word translation models and methods using Markov chains were introduced on document sets for query models estimation. The Markov chain method linked algorithms from link analysis/social networks. This was evaluated on TREC collections and compared to basic language modeling and vector space models along with query expansion using Rocchio. Major improvements were seen over regular
query expansion methods for strong baseline TF IDF systems, with highest improvement being for web data related short queries.

Long queries include extraneous terms hindering relevant documents retrieval. Kumaran and Carvalho (2009) suggested techniques to reduce long queries to effective short ones without extraneous terms. The work was encouraged by observing that totally lowering long TREC description queries lead to mean average precision improvements of 30%. The proposed approach necessitated converting reduction problem into a learning problem of ranking original query subsets (sub-queries) regarding their predicted quality, and selecting top sub query. Many query quality measures were described in literature as sub query representing features to train a classifier. Replacing original long query with top-ranked sub query selected by classifier results ranking lead to an average 8% improvement on test sets. Result analysis reveals that query reduction suits moderately-performing long queries, and query quality predictors small sets are compatible to rank sub-queries.

Berberich et al (2010) used a seamless approach to integrated temporal expressions in language modeling framework. Experiments reveal temporal expressions being useful in satisfying temporal information needs when their uncertainty is considered. Additional experiments/links for downloading extracted expressions and relevance assessments are seen in their technical report.

Lv and Zhai (2009) compared five representative state of the art methods to estimate query language models with pseudo feedback in ad hoc information retrieval. This included two variants of relevance language model, two variants of mixture feedback model, and divergence minimization estimation method. Experiments proved that relevance and mixture models variants outperformed others. Many heuristics were proposed intuitively
related to estimation method’s performance and reveal variations in heuristics implementation in varied methods to ensure an explanation of empirical observations.

Query likelihood retrieval was empirically effective for retrieval tasks. From a theoretical perspective, justification of conventional query likelihood retrieval function needs an unrealistic assumption ignoring a document’s “negative query” generation, suggesting it to be a non-optimal retrieval function. Lv and Zhai (2012) attempted improving query likelihood function by reverting negative query generation. A procedure to estimate negative query generation probabilities based on maximum entropy principle proposed derived complete query likelihood retrieval containing a negative query generation component. The proposed approach bridged theoretical gaps in existing query likelihood retrieval function and also improved retrieval effectiveness without additional computational cost.

Documents temporal aspects impact relevance for certain queries. Efron and Golovchinsky (2011) built modeling of temporal information on earlier work. An extension to Query Likelihood Model including query-specific information for estimating rate parameters introduced a temporal factor in language model smoothing and query expansion through use of pseudo-relevance feedback. These extensions were evaluated by using a Twitter corpus and two newspapers article collections. Results show that compared to earlier approaches, the models are better at capturing relevance’s temporal variability associated with certain topics.

Lafferty and Zhai (2003) provided a probabilistic semantics unified account underlying language modeling and information retrieval’s traditional probabilistic model revealing that both procedures can be seen as equivalent probabilistically, based on different factorizations of same generative relevance model. Discussions were on how both procedures lead to varying
retrieval frameworks in practice, as it involved different component models estimations.

Retrieval’s language modeling approach proved to perform well empirically. Its statistical foundations are this procedure’s advantage, but, feedback, as a retrieval system component was dealt with heuristically by the proposed approach. Original query was literally expanded by addition of more terms leading to expansion-based feedback creating an inconsistent original’s interpretation. Zhai and Lafferty (2001) presented a principled feedback approach in language modeling where feedback was treated as updating query language model due to availability of extra evidence in feedback documents. This model-based feedback strategy fit into a language modeling approach’s extension. The two different proposed approaches were updated using feedback documents based on query language model. One was based on a feedback documents generative probabilistic model, and the other on minimization of KL-divergence over feedback documents. Experiment showed that both approaches were effective, outperforming Rocchio feedback approach.

Jiang et al (2012) proposed/experimented on an adaptive browsing model. The authors wanted to present users with relevant and new results in multi-query search which discounted a document’s ranking in current search due to chances that it was clearly examined by users in earlier searches in same session. A document is penalized greatly when it appears in earlier searches or when ranked in higher positions in similar searches. Documents were also ranked by combining browsing model with ad hoc search models considering users’ browsing novelty in a multi-query search session. Experiments demonstrated the browsing model discounting documents found/ranked in higher positions in earlier searches, with relevant documents
being shuffled to higher positions to ensure that ad hoc search performance is unaffected.

Ponte (1998) suggested use of probabilistic language models to investigate text-based information retrieval and related issues. A problem was predicting text topic shifts, a topic segmentation issue. It showed that probabilistic methods can predict topic changes with regard to new event detection. Two complementary features sets were studied individually and combined into one language model. Language modeling allows a principled approach to this problem without complex semantic modeling. Next, document retrieval responding to user query was investigated. The proposed approach’s advantage was collection statistics used directly in language model probabilities estimation. The proposed approach performed well on query sets and standard test collections when tested. Empirical results provided more concepts to use language models for retrieval.

Wikipedia articles were used to semantically inform query models generation by Meij and Rijke (2010) for which supervised machine learning was used to automatically link queries to Wikipedia articles/sample terms from linked articles to re-estimate query model. On a big web corpus, substantial gains regarding traditional metrics and diversity measures were noticed.

Modeling query concepts via term dependencies ensured great positive effect on retrieval performance, especially in web search, where high rank relevance was critical. Earlier work treated concepts as equally important, an assumption that usually fails with regard to longer, and more complex queries. Bendersky, et al., (2010) showed that effective existing term dependence models can be extended by assigning weights to concepts. It was proved that weighted dependence model can be trained with existing learning-to-rank techniques, even with limited training queries. The study compares
effectiveness of both endogenous (collection based) and exogenous (external sources based) features to determine concept importance. To test weighted dependence model, experiments were undertaken on publicly available TREC corpora and a proprietary web corpus and results show the proposed model consistently/significantly outperformed both standard bag-of-words model and un-weighted term dependence model. It was seen that a combination of endogenous and exogenous features resulted in retrieval effectiveness.

Ostrogonac et al (2012) proposed a method of creating language models for highly inflective non-agglutinative languages. Three different types of language models were considered such as a common n-gram model, an n-gram model of lemmas and a class n-gram model. The last two types were specially designed for the Serbian language reflecting its unique grammar structure. All the language models has been trained on a collected data set which incorporates several literary styles and a great variety of domain-specific textual documents in Serbian. Language models of the three types were created for different sets of textual corpora and evaluated by perplexity values they have given on the test data.

Schulam et al (2012) presented the Source Code Statistical Language Model data analysis pattern. Statistical language model was an enabling tool for a wide array of important language technologies. Speech recognition, machine translation, and document summarization rely on statistical language models to assign probability estimates to natural language utterances or sentences. The process of building n-gram language models over software source files was described

Liu et al (2013) proposed a language model (LM) to handle the issue of poor context coverage in word sequence. In order to exploit the complementary characteristics of paraphrastic LMs and neural network LMs (NNLM), the combination between the two has been investigated. With the
use of a paraphrastic multi-level LM modeling both word and phrase sequences, significant error rate reductions of 0.9% absolute (9% relative) and 0.5% absolute (5% relative) were obtained over the baseline n-gram and NNLM systems respectively, after a combination with word and phrase level NNLMs.

Zhe et al (2013) proposed the temporal kernel neural network language model, a variant of models. This model explicitly captures long-term dependencies of words with the exponential kernel, where the memory of history was decayed exponentially. In addition, several sentences with variable lengths as a mini-batch were efficiently implemented for speeding up. The results were experimental shows that the proposed model was very competitive to the recurrent neural network language model and obtains the lower perplexity of 111.6 (more than 10% reduction) than the state-of-the-art results reported in the standard Penn Treebank Corpus.

2.6 CONCEPT-BASED APPROACHES IN INFORMATION RETRIEVAL

With the internet becoming faster, people tend to search and learn fragmented knowledge from it. Usually, vast documents, homepages or learning objects, are returned by powerful search engines in no specific order. Even if related, a user has to move forward/backward in the material trying to find which page to read first as the user might have limited experience in that specific domain, though a user may have domain intuition, which may in all probability be disconnected /disjointed. Hsieh, et al., (2008) suggested a learning path construction approach based on modified Term Frequency – Inverse Document Frequency (TF-IDF), Average Term Frequency – Inverse Document Frequency (ATF-IDF), and Formal Concept Analysis (FCA) algorithms. The approach first constructs a Concept Lattice with keywords from ATF-IDF collected documents to form a relationship hierarchy between
all keyword represented concepts. FCA then computes mutual relationships among documents to decide on a suitable learning path.

Kashyap et al (2011) presented BioNav system, a novel search interface enabling a user to navigate many query results by organizing them through the MeSH concept hierarchy. Biomedical databases search queries like PubMed return many results of which only a small subset is user relevant. Ranking and categorization combined were suggested to alleviate information overload. The focus of this work is results categorization for biomedical databases. MeSH annotations is a natural way to organize biomedical citations which is a concept hierarchy used by PubMed. First, query results are organized into a navigation tree with each node expansion step, where BioNav reveals a small concept nodes subset, selected to reduce user navigation cost. Previous works in contrast expand hierarchy in statically, without navigation cost modelling. It is shown that selecting best concepts reveal at each node expansion is NP-complete and suggests an efficient heuristic and an optimal algorithm for smaller trees. Experiments proved that BioNav outperformed state-of-the-art categorization systems greatly regarding user navigation cost.

Tsai et al (2006) proposed an adaptive personalized ranking mechanism to recommend SCORM-compliant learning objects from internet repositories. Digital courses have many learning units/learning objects, creating many learning objects according to SCORM standard. In the future, a huge number of SCORM-compliant learning objects will be published/distributed across the net. Facing this, learners will find it very hard to select suitable learning objects. The proposed mechanism uses preference-based and neighbour-interest-based approaches to rank the relevance degree of learning objects for a user. Through this, a tutoring system provides suitable learning objects easily for active learners.
Hubscher et al (2010) proposed a method allowing instructors to combine a general criterion and a context-specific preference flexible set to describe group types preferred. Group projects are important in many courses. This enables instructors to allow students to form own groups or assign them to one to increase group effectiveness. Computational tools supporting student’s assignments in class projects use general criterion, for instance, maximizing group member’s diversity. The instructor needs to consider frequently additional context-specific criteria/preferences which force him to figure out student’s assignments manually instead of through software. This task, though difficult and time consuming, results in suboptimal assignments. The proposed heuristic Tabu Search algorithm locates solutions that satisfy many preferences.

A current Web search problem is short and ambiguous search queries which are insufficient to specify precise user needs. To overcome this, search engines suggest terms semantically related to submitted queries to enable users to choose from suggestions that which reflects their information needs. Leung, et al., (2008) introduced an effective approach to capture user’s conceptual preferences to provide personalized query suggestions. This is achieved through two new strategies. First, online techniques that extract concepts from the web-snippets of the search result returned from a query are organized and concepts identify related queries for a specific query. Second, a new two phase personalized agglomerative clustering algorithm capable of generating personalized query clusters is proposed. No earlier work has addressed personalization for query suggestions. A Google middleware was developed to evaluate the technique’s effectiveness, and this was done through collecting click through data for evaluation. Experiments showed this approach to have better precision and recall than current query clustering methods.
Iosif et al (2010) presented Web-based metrics to compute words’ and terms semantic similarity and were compared with state of the art techniques. From the fundamental assumption that context similarity implies meaning similarity, relevant Web documents were downloaded through a Web search engine and contextual information of words of interest compared (context-based similarity metrics). The proposed algorithms work automatically, without human-annotated knowledge resources, e.g., ontology. It can be generalized /applied to differing languages. Charles-Miller data set and a medical term data set evaluated context based metrics revealing that context-based similarity metrics greatly outperformed co-occurrence-based metrics regarding human judgment correlation for both tasks. Additionally, the proposed unsupervised context-based similarity computation algorithms are competitive with state-of-the-art supervised semantic similarity algorithms using language-specific knowledge resources. Context-based metrics had correlation scores of up to 0.88 and 0.74 for Charles-Miller and medical data sets, respectively. The stop word filtering effect is investigated for word/term similarity computation. Performance of context-based term similarity metrics is evaluated as a web documents function for various feature weighting schemes.

Lee et al (2006) presented a semantic-aware learning object retrieval based ontological approach with two novel features: an automatic ontology-based query expansion algorithm to infer and aggregate user intentions based on original short query, and an “ambiguity removal” procedure to correct incorrect user queries. This approach can be embedded in other LOM-based search mechanisms to ensure semantic-aware learning object retrieval. The proposed approach - focused on digital learning material, contrasting other traditional keyword-based search technologies has demonstrated experimentally, greatly improved retrieval precision and recall rate.
Dilshener (2012) proposed an improved information retrieval based on the concept location by the contextual relationship. The existing technique was based on information retrieval (IR) which provides an adequate solution. Such techniques usually consider the conceptual relations based on lexical similarities during concept mapping. Proposed work uses the domain specific ontological relations during concept mapping and location activities when implementing business requirements.

Moon and Yoon (2013) presented a keyword base concept model for information retrieval. A keyword-based concept network was a method with the application of ontology. However, the proposed model has been added by association information between keyword concepts as a method for a user's efficient information retrieval. Also this concept network contains a keyword centered concept network, expert-group-recommended field concept network, and process concept network in it.

Shan et al (2012) proposed a meta search engine personalized mechanism which was based on ontology with the use of the Agent technology which mine the user behavior of characteristics with Agent for information retrieval by combining ontology technology set up and update the user general view which thereby reflects the user current interest and to the user's interest to the renewal of the dynamic. Based on ontology users, the user's query scene reasoning and filter search has been made. The results makes the search results to meet user needs, has the personalized.

Ta and Thi (2012) proposed an algorithm for improving the formal concept analysis to construct the ontology domain. With different purposes, both Domain Ontology and Formal Concept Analysis (FCA) were models used to present modeling concepts. FCA uses a lattice in mathematics to present concepts based on objects and attributes, where domain Ontology also presents the concepts and was used in various areas such as biology,
information retrieval, information extraction, etc. hence to build Domain Ontology based on FCA in order to support the information extraction task on a specified domain has been proposed.