CHAPTER 2

LITERATURE SURVEY

2.1 INTRODUCTION

With the enormous growth of internet and information available on the web, the IR process has become a fact of life for the internet users. Though internet provides vast and required knowledge to the users in all domains from different resources, some survey report directs that the users are not satisfied with the performance of the current generation of search engines for the reasons such as slow retrieval speed, communication delays and poor quality of retrieved results. With that concern, this chapter discusses the existing methods in the literature that are targeted to resolve some of the problems related to web-based IR.

2.2 TERM FREQUENCY-BASED RETRIEVAL

Troy et al (2007) proposed a relevance scoring scheme named Chronological Term Rank (CTR) based on the term frequency. CTR independently scores each term in the query with respect to the occurrences of the words in the document. The specific considerations focused by them are on multiplicative versus additive, absolute versus percentage, document length, log values, range variance and stop words. The authors have also claimed that the CTR outperforms the existing term frequency approaches to relevance scoring by enforcing fine-grained information regarding document structure in the relevancy determination process.
Muller et al (2010) described a method for term frequency-based IR. They have considered their work base as the bag-of-words based IR model. Additionally, the semantic information comprises the relatedness of query and document terms that are considered for effective results. Klein et al (2008) discussed the computation of term frequency and document frequency. Moreover, they have described computation time for Kendall in association with Web as Corpus (WaC).

Yamamoto et al (2001) presented an analytical model for effectively computing term frequency and document frequency for all terms in the document collection. They have incorporated suffix arrays for determining the frequency and location of the substrings that were grouped into a number of equivalent classes. While computing those frequencies, longest common prefixes, classes of substrings have also been considered. Specifically, this paper work concentrated on statistical language processing for bigrams and trigrams, working with much longer n-grams.

Term weighting for IR based on Linear Time Series Models has been proposed by Efron (2010). Here, the author has concentrated on determining the term collection frequency at discrete time intervals throughout the lifespan of the corpus. Under linear models for time series, he has considered moving average, simple linear regression and box-jenkins autoregressive models. Further, the term weights have been computed from the time series analysis.

### 2.3 SCORING USING TERM OCCURENCES

Similarity scoring using document ranking has been effectively accomplished by Anh et al (2005). They have established impact-based similarity scoring mechanism that is used on document at a time, and it
affords to fast ranking and high-level retrieval effectiveness. The paper analyzed the quantitative and qualitative impacts.

In a different way, fuzzy proximity degree of term occurrences was considered for appropriate IR form web by Beigbeder et al (2005). Boolean information retrieval has been enforced by the proximity operators which is a kind of AND. By analyzing different Boolean and extended Boolean models, the authors focused on developing local term proximity model, local query proximity model and global document query proximity model. Moreover, the position of the term occurrences in the documents is used for computing the relevance score.

2.4 DOCUMENT WEIGHT-BASED RETRIEVAL

A new term weighting model called Revision History Analysis was developed by Aji et al (2010). The revision history had been analyzed to redefine the term frequency. The document ranking has been performed here in association with two models, namely BM25 and the generative statistical lingual model. The historical revision of documents as a linear sequence of edits has been considered in the aspects of global revision history analysis, revision history burst analysis and edits history burst detection. In order to include structure weighting in document retrieval, Andrew (2005) proposed a process which deals with structured IR by corpus tree construction.

2.5 CONCERNING KEYWORDS

A methodology for effective value computation for keywords has been enforced in Byers et al (2010). Value per Click (VPC) estimation algorithm has been proposed to assign keywords or channels over time. The estimation algorithm is handled in three aspects.
- One keyword per measurement: Weighted Averages
- Multiple keywords per measurement: Linear Regression
- Strategies (oblivious strategies and adaptive strategies)

Bendersky et al (2008) evaluated and developed a technique that uses query-dependent, corpus-dependent and corpus-independent features for automatic key concepts extraction from verbose queries. The model incorporates ranking principle, concept identification and concept weighting. The concepts are classified on the basis of their weights.

2.6 INDEXING

Indexing plays a vital role in the process of IR in a facile manner. This has been analyzed by Salton et al (1975). The paper accomplished the document correlation effectively by the determination of similarity coefficient between the documents present in the defined document space. Clustering mechanism has also been enforced for appropriated indexing of similar documents.

Moreover, the geographic relevance had been well analyzed by Egnor (2010). In geographic information retrieval, the search results were effectively generated by indexing geographically relevant documents in accordance with the contents of the document and the multiple location identifiers. The process performed indexing by determining the location associated with the document, determining the plurality of locations associated with the document and the document reference.

Index extraction and index weighting have been studied by Jeong (2008). The steps given in the paper for document weighting is as follows: (1) Tag determination that will divide the paragraph, (2) Keyword extraction for
each tag, (3) Generating the weight vector for each tag, (4) Generating the document weight for each keyword, (5) Generating the vector value for user query, (6) Combining the weight vector and query vector, (7) Updating the document weight vector and (8) Restricting the index word weight

An alternative approach for indexing called IR-tree was enforced by Li et al (2011). The algorithm IR-tree in association with top-k document search algorithm concentrated on four major tasks, namely spatial filtering, textual filtering, relevance computation and document ranking in an integrated manner. Moreover, the documents are ranked on the basis of textual relevance and the spatial relevance. The performance evaluation of the concept had been made with storage overhead, index I/O cost etc.

Goerke et al (2011) presented a mechanism for retrieving the relevant documents for a set of documents based on the meaning of the query term and the search term input. The base for indexing was made by Sacks-Davis et al (1997) who enforced indexing document for queries based on structure, content and attributes of the documents. The indexing has been made in two aspects such as position-based indexing and path-based indexing. Moreover, the indexing attributes have also been well examined in this paper for proper indexing to get appropriate results.

2.7 HASHING IN RETRIEVAL

Hashing is a mechanism that supports well in indexing the relevant documents for effective information extraction process. With that concern, the paper proposed by Yao et al (2012) comprised a method for semi-supervised spectral hashing for adept similarity search process. Moreover, the paper considered the must-link and cannot-link patterns of supervised information amending the retrieval results in the aspect of labelled data and unlabeled data. In order to remove the binary constraints from the documents, the
hamming distance between terms has been measured with the square of Euclidean distance.

Following that, the recent work of Zhen et al (2013) provided an active hashing mechanism for efficient image and content retrieval from large repositories. The limitation of supervised hashing can be effectively processed by the active hashing mechanism, which involves in active selection of the most informative labelled pairs for hashing. The active selection process could be speeded up by a batch mode algorithm described in the paper.

2.8 DOCUMENT CLUSTERING TECHNIQUES

Document clustering is a powerful technique for large-scale topic discovery from text. Larsen et al (1999) proposed a method called Linear-time based document clustering for fast and effective text mining. The method comprises two steps: (1) Feature extraction maps each document and (2) Formation of hierarchy of clusters. The quality of cluster hierarchy is measured using F-measure. The clustering mechanism composed of processes such as seed selection, central adjustment and cluster refinement processes.

A comparative study was made by Steinbach et al (2000). Mainly, two document clustering processes such as agglomerative hierarchical clustering and k-Means were examined and studied. The authors have claimed that the hierarchical clustering process is limited because of its quadratic time complexity. On the other side, k-Means algorithm provides linear time complexity in the number of documents, but sometimes produced inferior clusters. Moreover, evaluation of the clustering technique has been accomplished with the parameters such as entropy, F-measure and overall similarity.
Law et al (2002) described a novel method for clustering called mixture-based clustering for feature selection. The significant task to be performed for feature selection process is the selection of data attributes that are most relevant for clustering. The paper incorporated the following two approaches (1) Estimation of feature saliencies based on Expectation-Maximization algorithm (EM) and (2) Extension of Koller and Sahami’s mutual-information based feature relevance criterion for feature selection.

A method for automatic subspace clustering of high dimensional data was presented by Agrawal et al (1998). They have presented a technique called CLIQUE that represents the dense clusters in the subspace of maximum dimensionality. The performance evaluation has been accomplished with synthetic data generation, synthetic data results, database size, dimensionality of the data space and dimensionality of hidden clusters.

A base for clustering searches and inverted index has been given by Voorhees (1986). The author has stated that the inverted index and clustered searches were highly responsible for returning a set of documents to the user with respect to the given query. The returned documents should have the highest similarities to the given query. He also analyzed the storage requirements and the processing time requirements for the process of effective indexing and clustering.

In an advanced manner, an approach for adaptive affinity propagation clustering was proposed by Wang et al (2007). This method outperformed the affinity propagation clustering that is having some limitations such as value determination for parameters and removal of oscillations. The method involved scanning the search space for a number of clusters and determining the optimal clustering solution. Cheng et al (1999) derived a clustering method for mining numerical data called Entropy-based subspace clustering. A database with numerical attributes was considered, by
which each transaction could be visualized as multi-dimensional vector. The authors have also described some of the parameters for subspace clustering such as criterion of high coverage, criterion of high density and correlation of dimensions. A different approach called cell-based clustering was proposed by Chang et al (2002). In order to perform the mining process in an efficient manner for large and high-dimensional data, this technique was enforced.

Liu et al (2000) derived a methodology for clustering based on decision tree construction. The decision tree has been used to partition the data space into clusters. Moreover, they defined that the clustering technique was developed based on the supervised learning method.

Lan et al (2009) derived a method for automatic text categorization. The term weighting methods are used for allotting appropriate weights to the terms to enhance the performance of text categorization. In this paper, there are three factors considered for text categorization such as term frequency factor, collection frequency factor and normalization factor. The authors compared the efficiency of their work with some benchmark methodologies such as combined term weighting methods, inductive learning methods, data corpora etc.

Phrase has been considered as an informative feature for enhancing the effectiveness of document retrieval process. With that concern, Hung et al (2008) proposed a novel technique for pairwise similarity computation termed as phrase-based document similarity. For that, they developed two models called Suffix Tree Document (STD) and Vector Space Document (VSD). The attained phrase-based similarity has been applied on the Hierarchical Agglomerative Clustering (HAC) algorithm. The typical definition for the phrase-based similarity is given as mapping up of each node labelled by a non empty phrase into a feature of M-dimensional term phrase. The vector space
document model gives the number of nodes in the node traversal and the various term frequencies.

A recent paper proposed by Li et al (2013) focused on multiple similarity measurement based heuristic hierarchical clustering. The authors termed the multiple similarities as the local alterations of initial document similarities based on the feedback information of the elementary clusters. The algorithm employed in the paper involves recursive modification of local similarities to attain efficient clustering. They used a clustering algorithm called Iterative Local Centroid Estimation algorithm. The multiple similarity measurement is achieved here in top-down process.

Another approach derived by Li et al (2008) explained the concept of text document clustering with respect to the word meaning sequences. Since the VSM treats the documents as bags of words, there is a possibility of ignoring word sequences. Addressing the problem, the authors proposed a novel clustering approach called Clustering based on Frequent Word Sequences and Clustering based on Frequent Word Meaning Sequences. The suffix tree has also been constructed here for effective retrieval of relevant documents in a facile manner.

2.9 TEXT CLUSTERING METHODS

In general, the text clustering methods are used to structure the large sets of texts or documents. Beil et al (2002) proposed a frequency term set based model for text clustering. Those frequent terms have been discovered using Association Rule Mining algorithm. For clustering, the measurement of mutual overlap of frequent term sets in accordance with the set of supporting document is required. Moreover, the authors framed two clustering patterns, namely flat clustering and hierarchical clustering. The paper considered the high dimensionality of data, large size of database and
understandable description of clusters. Apriori algorithm is used to determine
the frequent sets effectively.

The text clustering with feature selection using statistical data was
proposed by Yanjun et al (2008). The feature selection algorithm framed by
these authors is termed as CHIR and the text clustering algorithm is given as
Text Clustering with Feature Selection (TCFS). The algorithms concentrated
on $X^2$ statistics and new statistical data that can measure the positive term
category dependency.

Hotho et al (2003) explained the text clustering results using
semantic structures. Furthermore, the paper discussed on the way of
integrating large thesaurus and computation of lattices of resulting clusters
into common text clustering methods. The representation of text documents
was made with VSM that imposes the hierarchy of concepts in WordNet
thesaurus. Lattice was computed from the clustering results to correlate words
from the thesaurus hierarchy with different clusters and to compare the
different cluster representations. In the paper, the typical vector space
representation of text is extended by synsets of WordNet.

In text classification, feature clustering is a powerful technique for
reducing the dimensionality of feature vectors. Jung-Yi et al (2011) proposed
a fuzzy similarity-based self-constructing algorithm. Further, the words in the
feature vector of a document set are combined to form a cluster with respect
to the results of the similarity test. The approach can be applied in various
real time applications such as image segmentation, data sampling, fuzzy
modeling, web mining etc.

Concerning Text Clustering, a new feature selection method has
been proposed by Liu et al (2003). They named the algorithm as Term
Contribution and performed a comparative analysis on a variety of feature
selection methods of text clustering that include document frequency, term strength, entropy-based, information gain and $X^2$ statistics. In order to address the unavailability of label problem, the authors proposed an Iterative Feature Selection method that incorporated an effective feature selection method and clustering technique. The proposed method of this paper utilizes efficient supervised methods for increasing the clustering efficiency and performance.

A decentralized probabilistic text clustering was proposed by Papapetrou et al (2012). With this model, probabilistic model is utilized assigning document to clusters. The algorithm effectively performed the peer to peer clustering and enabled a peer to compare its documents with very few selected clusters without significant loss of clustering quality. The cluster indexing has also been performed for sorting the summaries of the clusters. The results are analyzed on the basis of quality, efficiency and scalability of text clustering in efficient IR.

In order to enhance text clustering, an efficient concept-based data mining model has been proposed by Shehata et al (2010). Statistical analysis has been performed for capturing the importance of the term within sentence, document and corpus levels. Additionally in this paper, the authors have also concentrated on verb argument structures, labels and concepts. The analytical framework has been given in the paper for computing the concept-based search with respect to the documents and sentences in a large corpus.

Fuzzy relational clustering algorithm for clustering the sentence level text was proposed by Skabar et al (2013). The algorithm generally operates on relational input data. Graph-based similarity and page rank has also been incorporated in this paper. The page rank score of an object is taken as a measure of its centrality of the cluster. The discussion part concentrated on the convergence, complexity, cluster duplication, thresholding of similar values and hard clustering.
2.10 RELEVANT DOCUMENT FILTERING ON RETRIEVING GEOGRAPHIC INFORMATION

A complete analysis of two approaches applied to the process of Geographic Information Retrieval (GIR) was given by García-Cumbreras et al (2009). The traditional query expansion process and the filtering process in efficient GIR were well analyzed by the authors in this paper.

Spatially - Aware Information Retrieval on the Internet (SPIRIT) was effectively described by Purves et al (2007). The authors concentrated on spatially-aware web search engine with a geographically intelligent user interface. The work used geographical and conceptual ontologies, term expansion and relevance ranking. The main components of SPIRIT architecture include query interfaces, geographic ontology, core search engine and relevance ranking.

2.11 CLASSIFICATION OF EFFICIENT CONTENT RETRIEVAL

In many applications Artificial Neural Networks have been utilized as an efficient tool for pattern classification problems. With that concern, the efficiency of Neural Network Classifiers was well analyzed by Sung-Bae (1997). The author has concentrated on three classifiers Multiple Multilayer Perception Classifier, Hidden Markov Model Classifier and Structure Adaptive Self Organizing Map classifier (SASOM). A description of Self Organizing Maps (SOM) was given by Kohonen (1990). It comprised four major components such as initialization, competition, co-operation and adaptation. The paper provided a detailed description of SOM algorithm for the effective clustering performance.

A PhraseFinder approach was proposed to construct collection dependent association thesauri and it was explained by Jing et al (1994). The
authors have also addressed some of the issues such as construction, access and evaluation to enhance the retrieval effectiveness. The text features that were used for framing PhraseFinder are as follows: terms, part-of-speech, paragraphs, sentences and phrases. Association generation and paragraph limits were used to represent the co-occurrences between terms and phrases within the text. Statistics and association data filtering were also considered for enhancing retrieval performance. Access to an association thesaurus had been determined with proper ranking methodology.

2.12 SEMANTIC INFORMATION RETRIEVAL

Markov et al (2006) proposed a paper on IR and Web Search that provided the description about two types of document collections such as traditional and web collections. It was stated in this paper that the index module is necessary for effective web search and IR. There are three kinds of indexing such as content index, inverter file structure based index and structured index. They have also described the ranking module, query module and page rank methods that are involved in the IR process. In order to enhance the process of traditional web searching, Semantic Search application was presented by Guha et al (2003). Moreover, the authors directed two Semantic Search Systems that are based on search query, and the enhancement of traditional search results with relevant data aggregated from different resources.

Sridevi et al (2011) presented a semantic similarity measure based clustering algorithm. The process performed document annotation followed by the clustering based on Particle Swarm Optimization clustering algorithm. SEAL approach was proposed by Maedche et al (2001). The model exploited semantics for construction, maintenance and the access to the information presented in the portal. Moreover, they have also described the ontology and
knowledge base and their differences. Semantic personalization has been well analyzed in this paper.

In order to retrieve the relevant image captions to the given query, Smeaton et al (1996) introduced a mechanism based on the computation of semantic distances between the words. Wei et al (2007) described an enhanced semantic-based retrieval approach. This framework has the description of network analysis, ontology-based information search and retrieval. The semantic association between two entities such as semantic web and IR was described. The work concentrated on the analysis of semantic web as well as indexing for effective retrieval.

An alternative approach called relation-based search engine in semantic web was proposed by Yufei et al (2007). Ontolook architecture was framed in this paper with respect to the microsemantic web environment. The Web Ontology Language (OWL) was well incorporated in this architecture. Further, a brief discussion was made by Mayfield et al (2003) about the integration of inference and IR on semantic web. In other words, it was stated that the framework supported both the retrieval-driven and information-driven processing. The retrieval process included the processing of semantic markup. Tokenization concept was utilized in which the words are considered as tokens.

The semantic search matching was naturally incorporated into the axiomatic retrieval model through defining the primitive weighting function based on the similarity functions of terms by Fang et al (2006). The authors have described some semantic term matching constraints. The performance evaluation had been made on the basis of the effectiveness of semantic matching and sensitivity.
Wasilewski (2011) explained the semantic query expansion process. The investigation was made with query expansion using a semantic modeling about the needed information. Furthermore, two ontologies are discussed, namely domain ontologies and automatic ontologies. The paper concentrated on four elements in every IR process such as analyzing the source of information needed by the user, user’s interface for entering the queries, inverting index of search engine and data repository for storing all collected documents.

The semantic web framework has been well described by Khatri et al (2012). The contribution of this paper focused on designing architecture for intelligent semantic web search engines. The paper also narrated some of the common issues in the existing search engines such as low precision and high recall, identity intention of the user and inaccurate queries.

Chiang et al (2001) presented a smart web query method for retrieving the semantic web data effectively. The method used domain semantics represented as context ontologies in order to formulate appropriate web query for web search. Moreover, semantic search filters have been utilized for identifying the most relevant documents and rank relevant web pages. The architecture of Semantic Web Query (SWQ) comprised an SWQ engine and its components such as query parser and an engine for context ontology determination. Semantic search filters are enforced for enhancing the precision-based properties in the context ontology.

An alternative approach called visual semantic modeling has been presented by Barb et al (2010). The model involved retrieving the relevant geospatial information with respect to content-based search. Association mining techniques are adopted for this model. In Natural Language Processing (NLP), different approaches have been proposed for quantification of lexical semantic relatedness. It was measured on the basis of WordNet
measures by Budanitsky et al (2006). The authors discussed the dictionary-based, thesauri-based, WordNet-based and other semantic network based methodologies. The approaches such as Sussna’s Depth-relative Scaling, Wu and Palmer’s Conceptual Similarity, Leacock and Chodorow’s Normalized Path Length for network scaling were discussed. Information-based and integrated approaches such as Resnik’s Information-based Approach, Jiang and Conrath’s Combined Approach and Lin’s Universal Similarity Measure have also been analyzed.

Furthermore, in the paper proposed by Du et al (2013) the authors have discussed the approach based on semantic ranking of web pages. The work started with the analysis of intension and extension similarity that noted the usage pattern and their hyperlinks. The semantic ranking given in the paper is stated in three aspects: (1) Hyperlink-based ranking, (2) Content-based ranking and (3) Hyperlink-content-based ranking.

The authors also included the construction of ISA and part-of-hierarchy using a user’s web log. Further, in the work of Fernández et al (2011), ontology-based semantically enhanced information retrieval approach has been proposed. Semantic indexing was carried out for effectively addressing the problem of knowledge extraction from text documents. Annotation has also been accomplished based on contextual semantic information. It was claimed in the paper that the annotation process used for semantic retrieval approach is very restrictive. Addressing that and to increase the annotation accuracy, the annotation process must be carried out when the documents contain the semantic context.

In a different way, the semantic similarity measures are used for enhancing the process of IR in folksonomies proposed by Uddin et al (2013). The annotation process of various web resources are based on the set of tags for effective searching purposes. The semantic similarity was computed based
on tag-user relationship, tag-resource relationship and WordNet. In order to determine the co-occurrences to select semantically relevant tags, the relationships between tag-user, tag-resources and tag-tag were analyzed.

2.13 ONTOLOGY DETERMINATION

Doan et al (2004) presented an Ontology Matching mechanism for finding the semantic mapping between two given ontologies. An architecture called GLUE was developed by the authors that employs the machine learning technique for creating semantic mapping in a semi-automatic manner. There is need of similarity measure in order to match the concepts between two taxonomies. With that note, the process utilized the concept of distribution-based similarity measures. Moreover, the authors also considered the type of learners to offer information. They have described learner types such as content learner, name learner and Meta learner. Multi learning strategy has been enforced for computing concept similarities.

A method called ORank was presented by Shamsfard et al (2006). Ranking the relevant documents according to the user query is required for efficient IR process. A new ontology-based ranking approach for sorting the HTML documents was proposed and evaluated in this paper. The approach combined the conceptual, statistical and linguistic methods for speedy retrieval of relevant information. The NLP techniques are incorporated for extracting phrases and stemming words. Following that, ontology-based conceptual method is used for document annotation and query expansion. Moreover, in this process both word and phrase extractions have been made for evaluating the relevance degree. In ORank system, the query processor is responsible for phrase extraction and words and phrases expansion. They have also stated that the proposed method enhances the precision rate of the retrieved documents.
A VSM has been derived for ontology-based information retrieval by Pablo et al (2007). The approach improves the search over large document repositories. Semi-automatic annotation of documents and a retrieval system has also been enforced in this scheme. Further, the semantic-based search is combined with the traditional keyword-based search methodology in addition to the annotation weighting algorithm for achieving tolerance to knowledge-based incompleteness.

From their method, the authors achieved better recall while querying for class instances, better precision using structured semantic queries, better precision by using query weights and better recall using class hierarchies and rules.

The process of ontology-based image retrieval has been well described in another paper presented by Vallet et al (2005). The proposed work included ontology-based scheme associated with the classic VSM, annotation weighting algorithm and ranking algorithm. Both knowledge base and document base are utilized in the approach. The authors have described the root ontology classes, document annotation methods, weighted annotation methods and ranking methods. The process inherited the well known problems of building and sharing up of well defined ontologies, populating the knowledge base and mapping keywords to concepts.

2.14 QUERY EXPANSION ON CONTENT MINING

An approach for Query Expansion (QE) in axiomatic model was described by Fang et al (2008). In this approach, the term similarity was determined based on the lexical resources. The term similarity exploited the information stored in two lexical resources such as WordNet and dependency-based thesaurus. Moreover, the following five relations were considered such
as (1) Synonym (Syn), (2) Hypernym (Hyper), (3) Hyponym (Hypo), (4) Holonym (Holo) and (5) Meronym (Mero).

The Concept-based QE proposed by Qiu et al (1993) was the base for incorporating the QE process in IR. A probabilistic QE model based on similarity thesaurus was constructed in the paper, where similarity thesaurus represents domain knowledge about some specific collections. The work concentrated on two significant issues with query expansion such as Selection of additional search terms and Weighting of additional search terms.

It is also to be stated that the approach relies on relevance feedback information only selected among the terms of a few retrieved documents.

2.15 DOCUMENT RANKING

A review work proposed by Larson et al (2004) reveals the evaluation results of algorithms of ranking methods in GIR system. Concerning the growth of digital spatial libraries, Beard et al (1997) proposed a multi-dimensional ranking for data. The ranking parameter based on keyword varies among search engines such as number of keywords presenting on the page, proximity of the keywords to each other, position of the keywords on the page and percentage of the keywords. The multi-dimensional ranking methodology that is in the paper composed of spatial, thematic ranking and temporal ranking.

SPIRIT provided by Joho et al (2004) presents an overview of large web collections. It was a funded project of EC fifth framework program with the motive of designing and implementing a search engine to find documents on the web with respect to regions or places referred in the given query.
Useful information about the GIR was given by Jones et al (2008). In that paper, the GIR was defined as the provision of facilities to extract and correlate rank documents or other resources from an unstructured or semi-structured collection on the basis of queries specifying both theme and geographic scope. With that analysis, the work of Perea-Ortega et al (2008) described GEOUJA system. The architecture comprised the following subsystems: Translation subsystem, Named entity recognizer subsystem, Geo-relation finder subsystem, Geo-relation validator subsystem, and Information retrieval subsystem.

A major confrontation regarding GIR is the quantification of spatial relevance of documents and generation of pertinent ranking of the results with respect to the spatial information required by the user. On addressing that, Cai et al (2011) presented a paper on relevance and ranking in GIR. Moreover, the document relevance is determined on the basis of the following parameters: Query-biased thematic relevance, Query-specific location relevance, Query-independent spatial relevance and User-centered conceptual relevance. These parameters are considered for finding the importance of the document in association with spatial and contextual significance.

The work described the importance of filtering for improving the Geographic Information Search. The experimentation was carried out with the GeoCLEF 2007 campaign. Moreover, the authors have tried to eliminate the retrieved documents from IR subsystem that do not satisfy certain validation rules.

Content similarity and novelty-based ranking have been enforced by Bordogna et al (2012). This work was presented on the consideration of “ranked list problem”. The three main phases of this paper are as follows: (1) The results obtained for the query is clustered, (2) Clusters are ranked-
based on personalized balance of the content similarity and query novelty and
(3) The disambiguated query from each cluster is highlighted.

Moreover the steps involved in accomplishing the tasks of the three phases given above are as follows: (1) Clustering, (2) Personalized novelty and similarity ranking, (3) Terms extraction from cluster candidates, (4) Disambiguated query generation, (5) Results visualization and (6) History updating.

Duh et al (2011) proposed an alternative method for ranking called semi-supervised ranking. The authors proposed a flexible transductive meta-algorithm that includes the key idea for adapting the training procedure to each test list after examining the documents that are required to be ranked. The main focus of this work is on the feature generation approach and the weighting approach. Instead of ranking the relevant documents, a methodology can be used for ranking the relevant documents for providing appropriate results to the user query. The work of Kayed et al (2010) described that ontologies can find concepts to aid machines to deal with data semantically. The first phase involves building the ontology concepts and the second phase involves measuring the relevance-based on the ontology concepts. Re-ranking has also been accomplished for enhancing the relevancy rate of the results.

Some of the recent descriptions on page ranking process in IR are given below:

For fast similarity search Lu et al (2013) proposed Rank Hash Similarity based approach that involves computing the top-k rank hash similarity for similar pairs and determining the low values. Moreover, Macdonald et al (2011) described the influence of document ranking in the web search methodologies. Derhami et al (2013) applied reinforcement
learning methodology on web page ranking algorithms to enhance the retrieval performance.

2.16 RETRIEVAL OF GEOGRAPHICAL INFORMATION

GeoVSM is an integrated geographical information retrieval model presented by Cai (2002) that combines the Geographical and VSM. The model comprised the document ranking by relevance accomplished by document-query similarity measure. The relevancy rate was computed on the basis of spatial domain and thematic domain.

The proposed architecture of this paper combined both the traditional GIS and traditional IR. In the work of Kavouras et al (2002) the formalization method for geographic categorization was discussed. The methodology is completely based on the theory of Formal Concept Analysis.

Conceptual integration-based semantic modeling was discussed and presented by Kuhn et al (2002). A classical integration example of houseboats and boathouses are analyzed. Moreover, the approach was tested with the functional language Haskell method. The authors have stated that the formalization of ontologies for geographical categories is carried out for two purposes such as (1) Creating basis for the implementation and usage of Geographic Information Service (GIS) and (2) Assessing the completeness and consistency of the ontologies.

An estimation algorithm for relevance computation for geospatial information called Geospatial Information Utility (GeoIU) was proposed by Meeks et al (2004). The method uses multi-attribute utility theory to access, score and weight the metadata queries that run against geospatial data and data acquired from distributed sources. Martins et al (2005) addressed the document ranking and indexing on geographical information retrieval system.
The general guidelines followed by the geo-retrieval algorithm are transformation of location and spatial operators in the query into geo-scope, Rank the geo-scope based on relevancy and acquiring the ranked list matching the set of geo-scopes. Lee et al (2000) presented their paper on Global Atlas that leverages on cartographic paradigm to afford a very natural support for indexing, searching and sharing of the information. They have used Generic Mapping Tool (GMT) and its database for multiple map generation of the same region with different projections.

Efficient storage and retrieval methodology for exploiting the geospatial and chronological characteristics was highlighted by the recent work of Malensek et al (2013). Clough et al (2005) demonstrated a method for extracting metadata in the process of spatially-aware information retrieval process. The work has been influenced with the following parameters: Speed, Reliability, Flexibility and Multilingualism. The main operations they have performed are Geo-parsing and Geo-coding. Then, the analysis has been carried out with the documents collected on SPIRIT. It is claimed in the paper that retrieval performance can be improved by combining various geographical resources.

2.17 OTHER RETREIVAL SYSTEM

Demonstration on Service Enhanced Retrieval System was given by Schaer et al (2010). It was developed on the basis of three model driven retrieval services such as Co-word analysis based query expansion, brodfordizing and author centrality.

The QE has been performed by the method of search term recommendations, brodfordizing and author centrality for re-ranking.
In general, based on the user query the current web search engines return a list of individual web pages ranked by their relevancy rate. It is to be noted that though the relevant information had spread across multiple pages, the process will retrieve an individual page. With that concern, the work of Varadarajan et al (2008) reveals a method for generating new pages called composed pages that contain all the relevant query terms. Ranking of the composed pages is performed on the basis of the hyperlinked structure of the original pages and the association between the keywords in each page. The approach is effectively applicable for long and multi-topic queries.

In a different way, SOM based on structure adaptive was well discussed by Kim et al (2004). It was pointed as SASOM. Moreover, fuzzy integral process adopted in this paper was for defining the relevance of classifiers. The method can be effectively employed for web content mining process for determining the user’s preferences as user profile.

The base for Geographical Information Retrieval was given by Agosti et al (1993). The authors prescribed a method for effectively managing the geographical and structured data associated with textual and other media data. Moreover, the work introduced an architecture and design approach for GIR system that is able to support and retrieve based on content and browsing on textual data. Types of media that are managed on the GIRS also described indexing that populates the semantic structure that combined with each document in the database.

Furthermore in this paper, three-level architecture was given for effective retrieval of content from the web. The representation and management of collection of interest was performed in the first level. The process of indexing has been defined and performed in the second level of architecture. The third level represented the plane of metadata classes.
2.18 SUMMARY

In this chapter, retrieval of information for a given query is outlined from the literature using the following techniques. The term frequency of the text and keywords comprises that are selected and some of the features that are extracted from the document. Documents are also assigned weights based on the scoring. The application of data mining techniques such as clustering techniques are studied on the text and documents by applying appropriate filters on the documents. Relevance of retrieval was also considered for improving the accuracy of retrieved documents. The need for the semantic retrieval and the presence of ontology in the documents and its impact in the retrieval are also outlined. Query expansion was described which widens the search space of the documents. The process of document retrieval is always a scope for further research and hence this thesis proposes document retrieval for the given query with new techniques.