CHAPTER 2

LITERATURE SURVEY

The literature survey has been carried out to explore the previous research works done in the following relevant areas:

- Web Mining
- Web Content Mining
- Extractive Summarization
- Query Based Summarization
- Sentence Scoring Approaches
- Preprocessing for Query Based Summarization
- Web page Filtering Approaches
- Web page Segmentation
- Applications of Summarization
- Applications of Query Based Summarization

2.1 WEB MINING

World Wide Web is accumulated with heterogeneous information contents to cater to the information needs of various user communities, and has now become the single largest collection of data in the information era. This enormous corpus which is actually growing beyond the expectations and imagination of everyone opens up scope for more research work.
Data mining techniques could be applied to this huge collection of diversified contents to uncover hidden patterns, trends and useful knowledge which could be utilized to provide value added services to web users. Few of the value added services resulting due to the knowledge gained are, search result ranking, targeted marketing, improved customer relationship and service, and fraudulent user or transaction detection. Therefore capturing and discovering knowledge from the web data resources has become very important for web mining research community.

Soumen Chakrabarti et al (1999) described a new hypertext resource discovery system known as focused crawler which analyzed its crawl boundary to find most relevant links and to avoid irrelevant regions of the web. This approach led to significant reduction in hardware and network resources requirement.

Mei Kobayashi and Koichi Takeda (2000) studied the growth and development of Internet and technologies that were useful for information retrieval on the web and also discussed the development of various techniques targeted to resolve problems such as slow retrieval speed and broken links associated with web based information retrieval.

Pre-processing of Web data to make it suitable for mining has been identified as one of the key issues for Web mining (Jaideep Srivastava et al 2002). An overview of the accomplishments of web mining and search engines (Jaideep Srivastava et al 2005) like ranking metrics, robot detection and filtering, information scent and user profiling were described and prominent web mining applications such as B2C E-Commerce applications like Amazon.com, visualization of the WWW, user tracking and personalized portals were elaborated.
Pranam Kolari and Anupam Joshi (2004), discussed various important research contributions in all three categories of web mining. In their discussion, it has been pointed out that web content and structure mining were being applied in areas like, ranking by contents, document classification, identification of hubs and authorities and web communities. It was also emphasized that web usage mining had paved ways for adaptive web sites, web page clustering and recommender systems.

Web usage mining techniques have been applied on web log data (Olfa Nasraoui et al 2008) for discovering and tracking evolving user profiles. Search engine has become the single most important application of Information Retrieval (IR) community (Bing Liu 2011) and the tremendous success of Web Search has pushed Information Retrieval to the central stage of WWW.

Web mining techniques have been applied to web contents, structure as well as web log data to improve the performance of web based services. Search engines, as the single most important application being used by huge number of different categories of users, have been attracting a lot of researchers to consistently improve the performance, user satisfaction and to provide additional value added services.

2.1.1 Web Content Mining

Web content includes text, hyperlinks, audio, video, images, graphs, animation, etc. Huge volume of heterogeneous web content provides vast scope for research community to innovate many ideas to improve the existing information services over web and also other added services. Generally, data mining techniques are applied on the contents present in web pages for automatic document classification, document clustering, finding the importance of a page with respect to a user given query, identify user interests
from user profile, identify advertisements that are relevant to the page content or query result, etc.

Cooley et al (1997) described various web content mining approaches like agent based and database based approaches and web usage mining tools for pattern discovery and pattern analysis. Raymond Kosala and Hendrik Blockeel (2000) discussed about research areas in web mining and different categories of web mining. Different research works done for unstructured data and semi-structured data from IR view have been summarized. In IR view, the unstructured text was represented by bag of words and semi-structured words were represented by HTML structure and hyperlink structure. In Database (DB) view, the mining always tried to infer the structure of the web site to transform a web site into a database.

Effective approaches for image retrieval from Internet which could also serve as an image search engine (Zheng Chen et al 2001) have been devised. Data mining techniques were applied on user feedback log file to improve the image retrieval performance. Useful knowledge patterns (Lizhen et al 2002) were captured from the web through searching, integrating, mining and analyzing. These knowledge patterns were used to build more efficient web sites to provide better service.

Web personalization system (Magdalini Eirinaki and Michalis Vazirgiannis 2003), emphasizing on web usage mining has been focusing on the process of customizing a web site to the needs of specific users. Knowledge acquired from the analysis of the user’s navigational behavior (usage data) in correlation with other information collected in the web context like, structure, content and user profile data have been utilized for this task.

Statistic and linguistic features of the document (Jaroslav Pokorny and Jozef Smizansky 2005) were used to determine a relevance rank of web
pages with respect to a given query. Sequence of text blocks were labeled as ‘Content’ or ‘Not Content’ (John Gibson et al 2007) according to the statistical machine learning method which were subsequently used to identify the required contents in the web document. Web content mining has been an efficient tool to extract structured and semi-structured contents from web and mining this content focused to obtain useful knowledge (Kshitija Pol et al 2008). Fast and efficient document clustering (Zhang 2006) using web mining approaches have been developed.

Web content mining techniques have been exploited in various ways to uncover hidden knowledge from the web page contents to meet the increasing user requirements and expectations. Web page ranking which is an application of web content mining plays an incredible role in achieving higher user satisfaction and in attracting more users. In spite of these, the search results provided by search engines contain more irrelevant, less relevant and redundant information content. This warrants for more time and effort of the user to correctly locate the required content even in the top ranked pages.

2.2 SUMMARIZATION

Summarization is the process of generating a small digest of the larger document which would help the user to have a quick preview of the document without reading the entire document. Summarization is an application within the information extraction domain where the required contents need to be identified, evaluated and extracted to provide the small digest.

The summarization process has three phases (Udo Hahn and Inderjeet Mani 2000): analyzing the source text, determining its salient points, and synthesizing an appropriate output. Summaries of threads of email conversations with automatically detected question-answer pairs
(Kathleen McKeown et al 2007) were devised using extractive summarization methods. Sentence similarity finder SimFinder (Hatzivassiloglou et al 2001) has been used to rate the similarity between sentences in email and the human generated summary and these summaries were then integrated as Question-Answer pairs of threads of emails.

Latent Semantic Indexing (LSI) has been utilized to measure sentence similarity (Hui Yu 2009) during sentence clustering from which news summaries were extracted. The concept of LSI was that terms representing sentence vectors were having some latent relations and the key component of LSI was Singular Value Decomposition (SVD) which reduced the dimensionality of vector. LSI considered the context of occurrences of two terms to measure the similarity. For example, the context of the words, “Internet” and “Web” would be useful to identify the common semantic meaning between them.

Correlation between ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores with human evaluation for extractive meeting summarization (Feifan Liu and Yang Liu 2010) was studied and meeting domain specific factors like filtering unimportant words that affect the correlation have been explored.

Multi document summarization (Xiao-Peng Yang and Xiao-Rong Liu 2008) system crawled through relevant web page, important sentences were extracted and re-organized to form a summary with the least redundancy. The most salient web pages within a set of topically relevant pages were considered for summarization but not according to user query.

Extractive summaries are created by extracting important sentences from the original content, based on statistical measures like TFxIDF, SimWithFirst, etc. Abstractive summary is created by rewriting sentences on
understanding the entire content of the original article by applying Natural Language Processing (NLP) techniques. The later technique is more computationally intensive for large data-sets.

2.2.1 Extractive Summarization

Summarization can be extractive or abstractive, in accordance with the nature of techniques applied to generate the summary. As the abstractive summarization is computationally intensive, the research community has been focusing on extractive summarization in which the document digest is built by extracting and stitching together the significant sentences from the actual content of the web page. Using supervised or unsupervised machine learning algorithms to extract sentences is currently the mainstream of the extractive summarization. Nowadays most of the supervised extractive methods deem summarization as a classification problem, and focus on finding good machine learning algorithms that can properly combine the features like proper names, sentence positions and sentence similarity etc.

Topic focused extractive summarization system (Lucy Vanderwende et al 2007) has been developed and the impact of various sentence simplifications and lexical expansion of topic words have been analyzed. WordNet synonym, morphological variants, geographical variants and word association data were considered for lexical expansion of topic words. The process of building this referential information was tedious and constructing an exhaustive word association collection was a challenge to automated, semi-automated as well as domain specific supervised systems.

Learning based approach (Kam-Fai Wong et al 2008) identified significant sentences using surface, content, relevance and event features of the sentences. Surface features like position, document _first, paragraph _first, length and quote, content features like centroid words, signature terms,
and high frequency words, event based score using page rank and relevance score using cosine similarity between related sentences were considered to measure the importance of a sentence to constitute generic summary of a document.

Multi document summarization system (Meng Wang et al 2009) in which the text document was divided into topic units has been built. These topic units were formed by considering the TF*IDF weight, event weight and topic weight of each word, and paragraph vectors were formed using these weights. Cosine similarity between the paragraphs was used to identify similar paragraphs to form the topic units of the document. Top scoring sentences were extracted from these topic units to compose the summary.

Individual summary was generated for each document in a set of documents selected for summarization and then common and different themes across these individual summaries were identified to cluster the summaries (Esau Villatoro-Tello et al 2009). Representative sentences from these clusters were extracted to construct the final summary of contents of multiple documents.

Distributed cluster summarization using key phrase extraction method (Khaled Hammouda and Mohamed Kamel 2009) was proposed, using intersection of documents that identified the candidate key phrases of a cluster. Upper level summaries in the hierarchical peer-to-peer document structure was obtained by merging cluster summaries of the lower level clusters.

Usage of Semantic Role Labeling (SRL) approach was exploited to score the sentences (Cem Aksoy et al 2009) instead of using individual terms for sentence scoring. SRL was expected to obtain more comprehensive information about the whole set of documents. Occurrence of frequent
semantic phrases in the sentences was used to score them and the top-scoring sentences were extracted to get a document digest.

Temporal multi document summarization system (Ruifang He et al 2010) extracted important sentences by treating temporally evolving information as dynamic objects. Important time points were identified over the macro importance discriminative model and summary sentences were extracted over these time points with micro importance discriminative model. Then both micro and macro importance discriminative models were integrated for summary content selection. List-wise learning-to-rank scheme (Xiaofeng Wu and Chengqing Zong 2010) has been utilized for supervised extractive summary generation.

Semi-supervised extractive summarization model based upon latent concept classification (Asli Celikyilmaz and Dilek Hakkani-Tur 2011) has been developed using a classifier which was trained on hidden concepts discovered from documents and their corresponding human generated summaries using a probabilistic Bayesian model.

Generic summaries are constructed by selecting the sentences that are closer to the main concept of the document. But user interest component is ignored which requires the additional effort of the user to locate the expected information. Information that is interested to the user but not closer to the main concept of document will not be considered for summary generation at all. This has the probability of reducing user satisfaction.

2.2.2 Query Based Summarization

In case of query based summarization, the sentence scoring would consider query relevant as well as query independent parameters to measure
its significance to include user interest component for producing the short extracts.

Use of Maximal Marginal Relevance (MMR), selected summary sentences relevant to user query and least similar to previously chosen sentences (Carbonell and Goldstein1998). Multi document and multi lingual summarizer MEAD (Dragomir Radev et al 2004a) utilized centroid feature, position, length, query overlap and keyword match feature of the sentences to identify the significance of a sentence for summary generation. SimWithFirst (Similarity With First Sentence) with MEAD (Md. Mohsin Ali et al 2009) has been used as a new feature to measure the significance of document contents.

Genre-independent multi document summarization (Toshihiko Sakurai and Akira Utsumi 2005) initially identified the document that was most relevant to user query. Core part of the summary was generated from this document and additional parts of the summary were generated from other relevant documents. All these summaries were merged to generate the informative summary for the user query. Weight of each sentence was calculated by adding the term weights of individual terms which was measured by means of term frequency and term co-occurrence frequency with the query. This system worked better for long extracts when compared to short extract.

Multiple documents summarization based on Hidden Markov Model (HMM) for sentence selection within a document and a pivoted QR algorithm to generate a multi-document summary (John Conroy et al 2005) has been developed. An HMM, in contrast to a naive Bayesian approach allowed the probability of a sentence to be in the summary to be dependent on whether the previous sentence has been selected for the summary or not.
Query relevant summary (Ramakrishna Varadarajan et al 2008) was composed by extracting and stitching together, the relevant contents from the URL selected from the search result and its hyper linked web pages. Distance between query terms (Hongguang Suo et al 2009) has been introduced as a new parameter in measuring the significance of each sentence in the document for automatic summarization. Summarization agent (Sun Park 2009) using Non-Negative Matrix Factorization (NMF) decomposed the matrix representation of a set of query relevant sentences into a non-negative semantic feature matrix and a non-negative semantic variable matrix. The objective function of this work was to minimize the Euclidean distance between the search result sentence matrix and its approximation. Generic relevance weight and topic score were used for generating the summary.

Sentence-based Latent Dirichlet Allocation (SLDA) for a new hierarchical model (Ying-Lang Chang and Jen-Tzung Chien 2009) was presented to characterize the structure of documents, sentences and words in a text corpus. The hierarchy of text corpus has been compactly represented by the associated sentence-based language models for the application of document summarization. Extracted sentences were arranged according to the structure of the document.

Various measures like surface measure, term based measure, event based measures and tag based measures have been utilized to score the significance of a sentence with respect to user query. Relevance score between the query string and the content will be able to capture the contents that are relevant but dissimilar to query string. For example, the words “internet” and “computer” are dissimilar but related to each other. Adding relevance score parameter to sentence scoring methodology will certainly improve the user satisfaction index of the summarizers.
Query based summaries are built at query time according to a query string given by the user. Entire contents of the document need to be processed to identify, select and extract the query relevant pieces of information to compose the summary which is a challenging task to the processing capacity and response time of the information servers. Some kind of preprocessing can be applied to reduce the run time overhead of the information servers.

2.2.3 Sentence Scoring Approaches

Comprehensive summary with respect to the given query, need to be generated by extracting important query relevant sentences of the documents. Importance of a sentence is measured by means of query dependent as well as query independent parameters of the sentences during query time.

Frequency of occurrences of query terms (Hui Yu 2009, Hongguang Suo et al 2009) in the sentence is an important factor in measuring the score of a sentence. As the frequency increases the score also will increase. Query terms may occur adjacent to each other in a sentence without any other word appearing in between or may not be adjacent to each other. As the average distance (Hongguang Suo et al 2009) between the query terms decreases the sentence score will increase.

Similarity measures like cosine similarity (Zhang Pei-ying and Li Cun-he 2009, Shuzhi Sam Ge et al 2011), Jaccard measure or Dice measure can be used to measure the relevance between the query and the sentence in the vector space. Semantic overlap (Wenjie Li et al 2005) between the query and sentence could be measured using lesk-similarity based on the word senses. Lesk-similarity does not utilize the senses it previously assigned and this greedy method is computationally intensive.
Position of the sentence (Hans van Halteren 2003, Kam-Fai Wong et al 2008, Hongguang Suo et al 2009) is considered to be an important factor in scoring the sentence. Normally, first and last sentences of a paragraph contain summary information about the paragraph and they are given higher score than sentences in the middle of the paragraph. First sentence of the paragraph is taken as a candidate sentence (Dragomir Radev et al 2004, Md. Mohsin Ali et al 2009) and all the sentences similar to the first sentence are given higher score.

Sentence features (Kam-Fai Wong et al 2008) based on content-bearing words i.e., centroid words, signature terms, and high frequency words are considered to measure the importance of a sentence. A centroid is a set of words that are statistically important to a cluster of documents. As such, centroids (Dragomir Radev et al 2004) could be used to classify relevant documents and to identify salient sentences in a cluster.

Phrasal words like, “firstly”, “therefore”, “to sum up”, “in one word”, etc., give a clue (Zhimin Chen and Jie Shen 2009) that the sentence contain summary information. Occurrence of these kinds of phrasal words in a sentence increases its candidature for final summary.

HTML tags like, <Title>, <Caption>, <Bold>, etc., try to give more emphasis to the enclosed contents (Zhimin Chen and Jie Shen 2009) compared to the adjacent contents. Occurrence of these tags will also improve the sentence score.

Salient sentences from the document contents need to be identified and extracted by considering query dependent as well as query independent parameters and the combination of various statistical and linguistic parameters produce different short extracts. Linguistic parameters consider only the semantic relevance which requires to be modulated with the
respective statistical scores of terms. A good mixture of linguistic, statistical and presentation parameters would certainly produce better summaries.

2.2.4 Preprocessing for Query Based Summarization

Query time overhead of the summarizers can be remarkably reduced by incorporating preprocessing phase prior to applying summarization algorithms.

Web page segmentation (Zhimin Chen and Jie Shen 2009) process has been included as a preliminary step at run time before applying summarization methodologies. DOM tree structure of the web pages has been utilized and similar adjacent leaf nodes were merged into topic blocks. Similarities between the leaf nodes were measured using cosine similarity. Query related topic blocks were processed to obtain the summary by considering the parameters like location of the sentence, enclosing HTML tag, prompt words like ‘sum up’, ‘in a word’, ‘to conclude’, etc. and frequency of occurrences of query terms in the sentences.

News summarization system utilized Sentence Similarity based Soft Clustering (SSSC) approach to split the document into clusters (Jun Tang and Xiaojuan Zhao 2009) and two different methods of sentence extraction were described. In first approach, sentences were selected from each of the query related clusters by turns, one after the other. In the second approach, query relevant clusters were merged into a single cluster and then top scoring sentences were extracted to build the short summary.

Document graphs (Shouyuan Chen et al 2009) were constructed where each vertex represented a sentence and edges were created according to the asymmetric relationship between sentences. Then the importance of a
subset of vertices was measured by adding a super-vertex into the original graph. The importance of such a super-vertex was measured which was a quantitative measure for the importance of a subset of vertices within the whole graph.

Web document has been mapped to a document graph (Furu Wei et al 2008) in which sentences were considered as vertices and semantically relevant sentences were connected through edges. Edge weight represented semantic closeness of the sentences connected by the edge. At query time, the sentence-to-sentence edge weight was calculated based on the given query and top scoring sentences were extracted to form the summary.

The above mentioned preprocessing approaches have been applied just before summarization during run time and building document graph becomes a runtime overhead for the system. These approaches have not focused on reducing processing overhead of summarization systems. Some preprocessing that can be carried out offline to improve the performance of the summarizer is required by reducing the size of the text unit which could be achieved by filtering the web pages contents.

2.3 WEB PAGE FILTERING APPROACHES

Task relevant contents of a web page could be automatically extracted with respect to a filtering criterion to reduce the textual unit size which in turn would reduce the processing time and response time of the automated task.

An automated tool to extract the user interested contents (Josep Silva 2009) from web pages has been built by traversing the DOM tree of the document, the nodes relevant to the filtering criterion were identified and
other irrelevant nodes were removed to obtain the user interested contents. This approach focused on providing the contents relevant to the user given criterion which may span many number of pages of contents.

Genre independent technique for detecting computer generated fake text (Thomas Lavergne et al 2010) in web pages using lexicographic features like mean and standard deviation of word and sentence length, ratio of grammatical words, etc was developed. Bayesian networks were utilized for information filtering system (Zhao Qingshan et al 2011) to filter related contents.

Two stage real-time event detection framework (Aixin Sun and Meishan Hu 2011) consisting of event fragment detection and event detection, identified event fragments according to the given query profile. A stream of documents were processed to identify event fragments by making use of Jensen-Shannon divergence metric and these fragments were clustered using the density based DBScan algorithm to obtain information related to an event.

These filtering approaches worked well in various scenarios mentioned above. DOM tree has been utilized in identifying user interested contents of a web page. This could be extended to fragment the page into set of conceptually related segments during pre-processing phase and used at query time for generating the short summaries.

2.4 WEB PAGE SEGMENTATION

A segment of the web page is a region that is identified as distinct from the rest of the page. Many web applications like information retrieval, information extraction and automatic page adaptation can benefit from the structure of the web page by decomposing it into several segments.
Text tiling using block comparison (Marti Hearst 1997) has been done to segment a web document in which adjacent pairs of text blocks were compared for overall lexical similarity. Text tiling algorithm subdivided text into multiple-paragraph units that represented passages or sub-topics for expository text that were not stylized or structured like environmental impact report without any sub headings.

Detecting the semantic content structure (Deng Cai et al 2003) of a web page potentially improved the performance of web information retrieval. User’s way of understanding the web layout structure was simulated based on visual perception. Spatial and visual cues could help the user to unconsciously divide the web page into several semantic parts. VIsion-based Page Segmentation (VIPS) algorithm aimed at extracting the semantic structure of a web page by making use of page layout feature. VIPS, first extracted all the suitable blocks from the HTML DOM tree, and then tried to find the separators between these extracted blocks. Separators denote the horizontal or vertical lines in a web page that visually crossed with no blocks. Finally, based on these separators, the semantic structure for the web page was constructed in a top-down approach.

Visual model of a document was utilized to segment (Daniel Egnor 2008) a web document. The visual model was determined according to the amount of visual white space or gaps within the document. Jorg Tiedemann and Jori Mur (2008) explored various segmentation approaches like co-reference chains, text tiling and window based methods and also concluded that fixed window based segmentation for passage retrieval performed better than other approaches.

Cohesive regions (Vineel 2009) of a web page among the unstructured contents having visually distinct segments, such as navigation
bars, advertisement banners, headers, etc were identified as different segments. Based on the characterization of DOM nodes, an unsupervised algorithm was developed to automatically identify segments of a given web page. A top-down algorithm (Ayelet Pnueli et al 2009) has been devised to obtain the layout of a web page from its rendered image. Layout algorithm was applied to divide the page into smaller and smaller components, according to the natural visual hierarchy to obtain meaningful segments of the page.

Semantically distinct regions (Yonatan Zunger 2011) of a document were determined by first translating the structured document into a hierarchical data structure containing set of nodes and edges. Nodes were annotated with a set of attributes including geometric parameters associated with the pseudo rendering of the structured document. According to the set of predefined rules, nodes were merged based on the attribute values like compatible attributes, nearby elements, etc.

A text stream has been segmented into documents (Jen-Tzung Chien and Chuang-Hua Chueh 2012) and in turn, into topic block using cosine measure to find the similarity between consequent sentences in the text stream and to find the document boundaries.

Many of the web page segmentation algorithms have employed visual rendering details as a hint for topic shift. But in practice, the conceptually similar contents need not be placed adjacent to each other. For example, the branch-wise sales details are placed under each branch but are not placed together. Cosine similarity measure which is based on bags of words concept, has been applied to measure similarity between contents to merge them for obtaining segments of a web page. This measure is less
efficient due to its sparse vector and also is able to capture only similarity but not relevancy between contents.

2.5 APPLICATIONS OF SUMMARIZATION

The application areas for automatic text summarization are extensive. As the amount of information on the Internet grows abundantly, it is difficult to select relevant information. Considerable amount of time and effort are required to browse, locate and extract the required pieces of information from the huge repository of information resources.

Automatic text summarization could automate the task of content reduction completely (Martin Hassel 2004) or at least assist in the process by producing a draft summary to tailor made, the newspaper content for different media channels like web newspaper, SMS messages, PDA, etc.

2.5.1 Applications of Query Based Summarization

Effective time management has become the critical need in the day to day life of everyone. Correctly locating the required information in the voluminous information repository is becoming a laborious task now.

Search engines present the snippets along with the URLs in the search results. This is expected to give a clue to the user about the usefulness of the particular document to the user. In general, snippets contain a set of non-coherent bits and pieces of sentences in which the query string occurs fully or partially. These snippets do not guide the users much in locating the expected information.

Multi document summarization has been proposed as an application to information retrieval. Important sentences from a set of topically related documents present in the search result were extracted and an indicative
summary (Dan Zhou and Lei Li 2007) was provided to give a hint about the content of the web documents in the result set which would guide the users for further reading of the entire content. Ranks of web pages in search result could be calculated by measuring relevance between the document represented by its summary and the query vector (Xian Jun Meng et al 2007).

Time complexity of document clustering could be reduced by clustering the respective topic based summaries (Maria Soledad Pera and Yiu-Kai Ng 2009) in which Naive Bayes classifier was trained on system generated summaries to classify the document collection. Word correlation factor was computed to identify representative sentences of a document. More number of applications were using summarization for the betterment of the text analysis and knowledge representation (Naresh Kumar Nagwani and Shrish Verma 2011).

Many of the above mentioned summarization systems were extractive in nature and were processing the entire document at query time. This affects the response time and processing efficiency of the system remarkably. Very few researches have been done to preprocess the web documents to improve the summarization efficiency. But those preprocessing are carried out as a preliminary step in summarization process.

Single and multi document summarization systems have focused on generating a digest of single or set of documents. Decision making and trend analysis process, requires analyzing the contents of a set of web pages with respect to the same query string. Summary of each of the document in the selected list need to be generated and these summaries need to be aggregated to have a comparative summary which will be useful for quick decision making.
This chapter presented a comprehensive survey on the web mining, web content mining, extractive summarization, query based summarization, preprocessing for query based summarization, web page segmentation and applications of query based summarization. This literature survey shows that query based extractive summarization has secured the attention of many researchers worldwide due to its usefulness to the information retrieval and information extraction domains.

The proposed frequent terms set and semantic relevance based segmentation algorithm is capable of segmenting the web document by grouping related contents within the document. The query relevant segments alone are processed at query time to generate the short summary. In the next step, comparative summary of a set of URLs selected by the user from the search result would be generated, which would help the user in content analysis, trend analysis and quick decision making. The objectives of research and the research methodology that is being followed in this research work are presented in the next chapter.