CHAPTER 1

INTRODUCTION

The advent of desktop systems in 1980’s and the growth of computer networks in 90’s lead to the use of computers and exponential increase in the size of data. Internet growth led to World Wide Web and today is the age of I-O-T. In the age of big data it’s processing has become a challenge today. Traditional data processing software is inadequate to deal with big data. Challenges include capture, storage, analysis, data curation, searching, transfer, visualization, querying, updating and information privacy. The database technology has replaced the traditional file processing systems. Today we talk of data warehouse and data mining techniques.

In current years the growth of the World Wide Web exceeded all expectations. World Wide Web is a huge, widely distributed and dynamic repository of data. The data size on the web is constantly increasing. The growth and diversity of electronic data available on the web has posed challenge for retrieving
relevant information from the web. Due to the exponential growth of the information on the Web and also the polysemous and synonymous characteristics of natural language relevant information retrieval on web domain is a challenge. The retrieved result set of documents to a query should be as far as possible accomplish users’ expectations. Mismatch and overload are two essential problems in efficient retrieval.

The literature survey in the area of web mining using traditional and ontology driven approach was done. Various literatures in the area of web mining and clustering methods were studied. Also a survey on layered architecture design was done. The literature survey in the area of web mining using traditional approaches and ontology driven approach was done.

To overcome the limitations of overload and mismatch various techniques are used by search engines and recommender systems. The search engines employ either supervised or unsupervised learning methods for information retrieval. Clustering of web documents is one of the methods that find a set of similar documents in a document corpus.

Classification is supervised learning method whereas clustering is unsupervised learning. In supervised learning, a model is learned using a set of fully labeled items. This set of fully labeled items is called the training set. Once a model is learned, it can be applied to a set of unlabeled items, called the test set, in order to automatically apply labels. On the other hand, unsupervised learning algorithms are completely based on unlabeled data. In unsupervised methods of classification the class labels are not predetermined.[1].
In unsupervised learning algorithms, the input data is not mapped to a predefined set of labels. Clustering algorithms take a set of unlabeled data as input and then group the items using some notion of similarity.

Classification is the process of automatically applying labels to data, such as emails, web pages, or images. Clustering is defined as the task of grouping related items together. The objective of clustering is descriptive and classification is predictive. The goal of clustering is to discover a new set of categories. The new groups are of interest in themselves, and their assessment is intrinsic. In classification tasks, the groups must reflect some reference set of classes and the assessment is extrinsic.

Clustering algorithms are broadly classified as hierarchical and partitional methods. K-means is a partitional algorithm. K-means uses minimum distance rule to arrange data items into different centroids. It compares the distances between a data item and each of the centroids and assigns the data item to the cluster with nearest centroid. Various distance measures such as Euclidean distance, Manhattan distance, and cosine similarity are employed for computing the distance.

Traditional clustering algorithms rely on Bag-on-Words approach. Bag of Words approach do not consider the semantic relationships between words. To overcome this limitation, introduction of semantic information from ontology such as Wordnet is used to improve the result of clustering process[2,3].

Hotho et al. applied ontology in document clustering for word sense disambiguation to improve the clustering performance. To
evaluate the performances three methods were considered: COSA, SiVeR and TES. COSA was evaluated against the elementary methods of SiVeR and TES. SiVeR or Simple Vector representation failed from the beginning. This is because clustering does not yield good results in high dimensional space. Thus another pre-processing method TES was considered. TES or Term Selection approach is based on the feature vectors but focuses on few terms. It produces a low dimensional representation. The selection of terms is based on the information retrieval measure tf/idf. Tf/idf measure ranks the terms on the basis of term frequency and inverse document frequency. For example, considering this measure terms that appear too frequently in a document are ranked lower as compared to other terms and do not contribute to the clustering results. But Term Selection method does not consider any background knowledge for resolving ambiguity in natural languages. The proposed approach COSA or Concept Selection and Aggregation overcomes this limitation by using background knowledge available in the ontology. Hotho et al. employed the concept of ontology for document clustering. They used pre-processing method, COSA. COSA pre-processing method maps terms to concepts and select between aggregations navigating top-down hierarchy. They used COSA as simple core ontology for restricting the set of relevant document features and for automatically proposing good aggregations. The core ontology defines the background knowledge used for pre-processing and selection of relevant views (i.e. aggregations) onto the set of texts. The actual ontology used is domain-specific. The ontology is manually modelled and describes tourism domain.
The performance of the three pre-processing methods was evaluated using K-means algorithm and the evaluation metric silhouette coefficient. The result of evaluation depicted that K-means based on COSA excelled as compared to K-means based on TES. Also K-means based on SiVeR could not result in reasonable clustering structures and the silhouette coefficient approached to value 0[2].

U. K. Sridevi and Nagaveni N. used KIM ontology to evaluate the performance of an information retrieval system using optimization algorithm. The experiments were performed using GATE Tool API and the dataset used is Newsgroup. In GATE information retrieval documents can be retrieved on the basis of their features or annotations. The pre-processing steps of stemming and stop word removal were performed.

The comparison of precision, recall and F-measure values depict that ontology similarity measure performs better than keyword similarity measure. F-measure values with ontology similarity were better than keyword similarity[3].

The domain of web mining lacks a unified and universal framework. There is no proposed architecture for the domain of web mining. The proposed layered architecture is an attempt to overcome this limitation. The layered architecture supports semantic clustering of web documents. Semantic text document clustering is implemented by using WordNet ontology in the preprocessing phase of document clustering.

The layered architecture provides a unified framework for clustering. The pre-processing function with WordNet ontology
is unified with traditional clustering methods to overcome the limitation of ambiguity in natural languages. The proposed unified theoretical framework is based on the following assumptions which are the phases of efficient clustering of document corpus:

Step 1: Select a document corpus that is candidate for clustering.

Step 2: Apply document pre-processing or document transformation steps of stemming and stop word elimination. For semantic pre-processing inclusion of WordNet ontology is performed before clustering.

Step 3: Document Clustering is performed in this step. The goal of clustering is to find groups or classes of objects that are very different from each other and the properties of members within a class or group are very similar.

Step 4: In this phase evaluation and visualization of results is performed. Evaluate the performance of clustering and plot graphs for the same is the objective of this phase.
1.1 Context and Background

World Wide Web is the single largest data source in the world. Web information is large and diverse. But only a small subset of the information is relevant or useful to a user. It is a challenge to find high-quality web pages on a specified topic. To retrieve relevant and useful information from huge web content is a difficult and challenging task. It is difficult for users to search and retrieve documents that are relevant to their particular needs. Users browse through a large hierarchy of concepts to find the relevant information. The query submitted to a search engine has to wade through irrelevant documents.

Web mining includes search and retrieval of information from resources such as Web documents. It is the discovery of knowledge from Web data. Data on the web includes content of Web pages, information available via hyperlinks and Web log data. It is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. Data mining refers to extracting or mining knowledge from huge amount of data.

Data mining is an essential step in knowledge discovery process [4]. The knowledge discovery process is depicted in Figure 1.
The objective of mining is to apply various methods and algorithms for extracting data patterns. Data patterns can be mined from various types of data repositories such as relational databases, data warehouses, transactional databases, object oriented databases, legacy databases, spatial and time related databases and World Wide Web. The data patterns mined represent knowledge related to particular domain. To mine data various data mining functionalities such as classification, clustering, and association rule mining are applied on data. A data mining task is formulated in form of a query that is input to the data mining system. The task or query is based on the following primitives[5]:

- Task relevant data- This specifies the data on which mining is to be performed. This includes databases, data warehouse and data repositories.
• Kind of knowledge to be mined- This includes the kind of knowledge to be mined. It determines the data mining function such as clustering to be performed.

• Background knowledge- It is information about the domain to be mined that is useful in discovery process.

• Interestingness measures- These define functions that evaluate the mining process. They separate non-relevant pattern from meaningful and useful patterns.

• Presentation and visualization of discovered patterns- Various different forms of knowledge representation such as rules, tables, graphs, charts are used for displaying the evaluated mining results.
The task of web mining is classified as [4,5,6]:

I. Web Content Mining
Web content mining is the process of extracting useful information from the content of web pages. The technologies used are Natural Language Processing and Information retrieval. Information retrieval is the process of organizing, storing, retrieval and evaluation of information relevant to user’s query. The retrieval is on basis of content and the output is a set of documents that contain information related to user’s query.

II. Web Structure Mining
The objective of structure mining is to categorise the Web pages and generating information such as the similarity and relationship between them. The hyperlink structure and topology of the web is major input for structure mining.

III. Web Usage Mining
The goal of web usage mining is to identify browsing patterns by the analysis of web server logs.
Web Content mining includes many concepts of traditional text mining techniques [7]. Clustering is one such method. Clustering aims at grouping data into clusters or classes. A cluster is a group of similar objects. The objects within a cluster have high similarity whereas objects in different clusters are highly dissimilar. Dissimilarity is evaluated on the basis of various attribute values that describe the objects. In general distance measures such as Minkowski distance, Euclidean and Manhattan are used for evaluating the dissimilarity between objects. Clustering algorithms require either a data matrix or a dissimilarity matrix as input. A data matrix or object-by-variable data structure represents n objects with p variables. The structure is in form of a relational table or n-by-p matrix. It is two-mode matrix as rows and columns represent different entities. A dissimilarity matrix or object-by-object matrix stores a collection of proximities for all pairs of n objects. It is represented by n-by-n table. The dissimilarity is measured as a difference between objects i and j. It is one mode matrix. Clustering algorithms are broadly classified as: Clustering can be broadly classified as:

- Flat
- Hierarchical
Flat clustering results in a flat set of clusters without any explicit structure that would relate clusters to each other. On the other hand, in hierarchical clustering a hierarchy of clusters is generated.

Hierarchical clustering can further be classified as agglomerative or divisive[8]. An agglomerative method starts with each document representing a single cluster. This is a bottom up approach to clustering. In divisive approach all observations start in one cluster and continue until each object is in a separate group. It is a top down approach to clustering.

Semantic clustering includes the semantic relationships between words. This enhances the meaning associated and results in better clustering of web documents. The traditional clustering methods rely only on Bag of Words approach and do not consider the semantics associated. This limitation of traditional approaches poses the challenges in effective clustering and mining relevant documents from document corpus. Semantic similarity imposed by semantic clustering methods overcomes the limitation of traditional methods. Semantic clustering is supported by including ontology as background knowledge. Mining with ontology yields better results as compared to traditional Bag of Words approach[9]. The inclusion of WordNet ontology enhances the semantic content and yields improved clustering results. WordNet is used for disambiguating terms and identifying concepts pertaining to a term. The identified concepts satisfy the associated semantic meaning.
The semantic relations of WordNet hierarchy expand the users’ query by the inclusion of synonyms. The semantic information of WordNet is valuable in tasks of document categorization and summarization.

Document clustering is the process of collecting Web sources such as HTML pages into groups so that similar objects are in the same group and dissimilar objects are in different groups. It is a method for classifying information into manageable meaningful classes or groups. It is a data reduction tool. It creates subgroups that are more manageable than individual datum. The goal is to maximize the dissimilarity between groups that are initially unknown. It creates new groupings such that each cluster are similar in some ways to each other and dissimilar to those in other clusters. A cluster is a group of relatively homogeneous observations or classes [10].

Clustering on the Web has been proposed based on the idea of identifying homogeneous groups of web documents. Clustering web documents comprises of two essential phases:

- **Preprocessing**

  It is initial step that prepares the document corpus for the next phase of grouping similar documents in a cluster. The step includes the following:

  *Stopword elimination*: Stopwords are words that have little semantic value associated but their frequency of occurrence in a document is high. These include prepositions.
**Stemming:** It is a process of linguistic normalization. In this process the inflated forms of a word are reduced to stem.

The variant forms of a word are reduced to a common form. For example "stems", "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument" [11].

Clustering is performed on data. The input data is represented as vectors. To represent text or input documents vector space model is typically used. Vector space model was given by Gerard Salton. The model represents the document set as an m×n term-document matrix. Each column represents a document. Each (i, j)th entry in the vector represents a weighted frequency of term i in document j.

In vector space model a document is represented as a vector of the terms that appear in all the document set. Each term in a document becomes a feature dimension. Each feature vector contains term weights of the terms appearing in that document. The term weighting scheme is based on tf.idf method. Tf.idf combines local and global statistics for evaluation. Tf or term frequency is the local component and
is a measure of the importance of a term in a particular document. The more a term is present in a document, the more that document represents that particular concept. Idf is the global component and assigns more importance to more specific terms. It emphasizes the terms that discriminate a document from other documents in the corpus [11]. Various algorithms can be applied on the preprocessed corpus to achieve the results. Clustering algorithms can be categorized as: hierarchical and partitioning methods. It is a method of clustering in which a set of data objects such as documents are grouped into a tree of clusters. A hierarchy of clusters is created and objects can be assigned to different numbers of clusters. The result is a hierarchy or set of nested partitions. A hierarchical clustering method works by grouping data objects into a tree of clusters. A hierarchical structure provides efficient access to information contained in a collection of documents[12]. These methods can further be classified into agglomerative and divisive hierarchical clustering depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Partitioning methods of clustering constructs various partitions and evaluates them by some criterion. A data repository of n objects and k is the number of output clusters, a partitioning algorithm organizes the objects into k partitions or clusters and \( k \leq n \). The clusters are formed to optimize an objective partitioning criterion. The objective partitioning criterion is a distance function such as Manhattan. The
The distance functions such as Manhattan between the objects is defined as:

\[ d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|. \]

K-means is most well known partitioning algorithm.

The focus is to retrieve the most useful and relevant set of documents from World Wide Web (WWW). Web is dynamic and information on the web keeps on updating and changing. Also new information is added and size of web increases. The growth of Web has caused of a number of problems with its usage. In particular, the quality of Web search and corresponding interpretation of search results are not according to users’ expectations. The retrieved information has drawbacks of mismatch and overload. Mismatch and overload retrieve irrelevant documents that do not match the user’s need. Mismatch implies some useful and interesting data is missed. Overload means some retrieved data is not what users’ require. Traditional methods of information retrieval face these challenges. The ambiguity in data terms is the main reason for the above problem. An ontology mining technique provides solution to the problem.

The objective of ontology mining technique is to automatically discover ontologies from data for building concept models of users’ need. The results show the relevance of ontology and the ontology model was successful. [13,14]
It is difficult for machines to understand the semantics and meaning associated with the content. The content lacks context and leads to various limitations of keyword based search and information retrieval. The input query is a string comprising of keywords. Also a document is represented as set of keywords or index terms. But same word can have multiple meanings in different context. This ambiguity needs to be resolved. The word ‘bank’ has several meanings such as an institution or river bank. The machine needs to resolve the context and find the most appropriate set of documents. The machine applies content mining for retrieving and finding from huge repositories of data.

The challenge has promoted to find methods for effective and efficient searching on the web. The challenge can be overcome by Web content mining. It aims at gathering information from web sources for knowledge discovery. One solution is to construct meaningful organizations of objects. The essential application is to group similar objects into classes. Classes or clusters are collections of objects whose intra-class similarity is high and inter-class similarity is low.

The traditional approaches to clustering do not capture the semantic structure of terms in a document. The clustering methods are ineffective when terms are used for cluster analysis. The results of clustering can be improved by inclusion of semantic structure associated with terms [15,16].
Keyword base search is unable to satisfy user’s need and insufficient for Web resource discovery. A huge set of documents are retrieved but only a small subset is relevant and useful. The problems of synonymy and polysemy lead to extraction of documents that are not specific to user’s need. The keywords entered by users may imply different information needs of different users. The limitation holds prominent in cases such as:

Polysemy: “jaguar” as animal vs. “jaguar” as car
Synonymy: “movies” vs. “films”.

This causes ambiguity during query processing and leads to unwanted results. The search quality is not as per users’ expectations. Thus information retrieval should not only focus on keywords but also on semantic and context associated. This improves the retrieval result and improves result returned for a user’s query.

Keyword based search ignores the semantic and contextual information in retrieval process. Semantic relations such as synonymy and polysemy are ignored. This limitation results in problem of mismatch and overload in resultant documents. Semantics is associated with meaning of the content [17].

For associating semantics, ontology is a major affix to the already existing documents. Ontology implies a set of concepts and the relationships that exist between the concepts. Ontology is an explicit specification of a set of objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold them. Thus ontology is a set of concepts $C$ and relationships $R$. The relationships in $R$ can be taxonomic or non-taxonomic. The instances of an ontology refer
to the instances of its concepts and relationships [18]. If each concept instance exists in the form of a Web page, a relationship instance will then exist in the form of a Web page pair. This view has been adopted in most of the Web classification research.

Ontology describes the information within an application domain. When ontology is used in solving a Web classification or extraction problem, the results obtained can be associated with the ontology entities making them easier to understand. For example, within the University domain {Professor, Student, Course} and {Teach, Register, Supervise} are the common concepts and relationships respectively[19,20]. University Web pages are likely to centered around these concepts and related concept instances are likely to be linked in one way or another. *Semantic clustering is supported by including ontology as background knowledge.*

WordNet ontology is one of the most widely used and largest lexical databases of English. WordNet as a dictionary covers some terms from every subject related to their terms. It maps all the stemmed words from the standard documents into their specific lexical categories. It groups nouns, verbs, adverbs and adjectives into sets of synonyms called synsets. The synsets are organized into senses giving thus the synonyms of each word and also into hyponym/hypernym(i.e. IS-A) and meronym/holonym(i.e. PART-OF) relationships
1.2 Motivation
Keyword based search do not consider the semantic relationships among words and thus the meaning of documents cannot be accurately represented. The traditional clustering methods do not consider the semantic relationships among words. The synonymy and polysemy problems are a major challenge for effective retrieval of web documents. To resolve the problem of word sense disambiguation and to retrieve the most relevant documents as per user’s query semantic clustering algorithms are proposed. Word sense disambiguation is the process of replacing original terms in a document by the most appropriate sense as dictated by the surrounding context of a document. Polysemy and synonymy are two essential problems that affect text representation and text clustering. The use of ontology as background knowledge is one of the solutions in semantic clustering algorithms. A wide range of semantic clustering methods are proposed for effective retrieval of web documents from document corpus.

Another issue in web mining is that there is no proposed architecture for web mining. The domain of web mining lacks a unified and universal framework. The framework can be provided by an architecture that maps to a pattern. The architecture pattern provides a solution to frequently occurring same or similar problem domains. It represents a structural organisation schema for a domain. It defines the various modules and roles of each module.

The layered architecture provides a solution to the stated limitation. Layered architecture lays emphasis on grouping of
related functionality within an application into distinct layers. The layers are stacked vertically on top of each[21,22]. The emphasis is to propose a universal architecture for semantic clustering of documents.

A wide range of semantic clustering methods have been proposed for effective retrieval of web documents from document corpus. But there is a lack of a universal architecture for semantic clustering of documents. The main objective of the dissertation is to provide a universal and unified framework of semantic clustering of documents.
The proposed multilayered architecture comprises of the following [23]:

**Layer 0:** It represents the Web itself. It is a huge and dynamic repository of information. It is widely distributed and contains a huge and rich collection of useful and relevant information. It contains hundreds of terabytes of information and still growing rapidly. It is merely impossible to set up a data warehouse to store and integrate all data on the Web. The web data is continuously updated and refreshed such as information in news, stock markets, and advertisements. The data on the web is semi-structured. It is not structured as in databases. Web pages lack a unifying structure and a predefined schema. Due to the implicit characteristics of WWW it poses challenges for effective and useful knowledge discovery.

**Layer 1:** It is the Web page descriptor layer. It contains descriptive information for pages on the Web. It contains information related to Web documents such as address of the page, keywords, timestamp; access frequency. The attributes are useful in web mining. The thesis focuses on keywords for web content mining.

**Layer 2:** The layer represents the clustering process. Cluster analysis is a technique for multivariate analysis that assigns items to automatically created groups based on a calculation
of the degree of association between items and groups. It deals with the organization of a set of objects in a multidimensional space into cohesive groups, called clusters. The groups or clusters which are formed should have a high degree of association between members of the same group and low degree between members of different groups. It is a tool of information discovery. It has the potential to reveal previously undetected relationships between data. The goal of clustering is to create classes or clusters such that the objects within a group are similar and related. The greater the similarity the better and distinct is cluster analysis. In context of web mining cluster analysis aims at grouping web pages on the basis of document content. It is useful for organizing documents to improve retrieval and support browsing.

![Layered Architecture](image)

Figure 1.2: Layered Architecture
1.3 Objective of the dissertation

This dissertation focuses on ontology driven web mining by clustering web pages. Web mining is the discovery and analysis of useful information from World Wide Web. The dissertation lays emphasis on web content mining. The information retrieval process focuses on web document clustering. The retrieval process is based on the content of web pages. The document clustering methods identify inherent classes of similar pages to a particular domain or topic. The output of the retrieval process is a set of documents that contain information related to user’s query.

The thesis lays emphasis on the role of semantics and ontology in web mining. The inclusion of ontology enhances the semantic and context and improves the retrieval process on the World Wide Web[24]. The focus is to retrieve and find the most useful and relevant. The challenge is to overcome the limitations of traditional keyword based search. A semantic search overcomes the drawbacks of overload and mismatch associated with keyword based search. The inclusion of semantic relationships among words improves the result of clustering. From a collection of web pages the aim is to retrieve the best relevant document to particular domain. Each class forms a cluster of similar web documents. Relevance denotes how well a retrieved document or set of documents meets the information need of the user. A document belongs to a class only if the relevance value is greater than or equal to a
specified threshold. Various metrics such as silhouette coefficient is used to evaluate the result of clustering process.

The emphasis is on semantic clustering of web documents using ontology as background knowledge. The inclusion of WordNet ontology enhances semantics and improves the clustering of documents.

The use of ontology to replace the original terms in a document by the most appropriate ontology concept improves the clustering result and improves the retrieval process. To overcome the flaw of Bag of Words approach and traditional clustering methods, a semantic approach to mining is required. A semantic approach is based on the semantics or meaning associated with units larger than words such as sentences. This is because words can be disambiguated in relation to the context. Word sense disambiguation requires semantic level processing. The advantages of associating semantics in clustering are:

- Resolution of synonym and polysemy problems in natural languages.
- Extracting core semantics from text.
- Assigning distinguished and meaningful description for generated clusters.

In this thesis enhancement to traditional clustering methods is presented by the inclusion of semantics via ontology. The objective is to improve the generated clusters quality by considering background knowledge via ontology.
The proposed approach to ontology driven mining and semantic clustering is performed by preprocessing the input document corpus with WordNet ontology. A Natural Language Processing language such as Python is employed for pre-processing the document corpus with WordNet. After pre-processing the document corpus clustering is performed. This requires the document to be converted into term-document matrix. Term-document matrix is the most common format to represent texts for computation. Then standard techniques from statistics and data mining, like clustering or classification methods can be applied. In my work, a traditional method of clustering such as K-means is applied to evaluate the result of clustering.

A comparative analysis of clustering result is performed with and without WordNet to interpret the change in clustering quality with the inclusion of WordNet pre-processing. To measure the cluster quality a metric such as Silhouette coefficient is required.

Another major objective is to provide a unified and universal framework for semantic clustering of documents. The framework proposed is multilayered architecture(Figure 1.1).The architecture maps to Layered Architectural pattern.
Based on the literature survey the contributions of the dissertation are as follows:

- To propose a multilayer architecture for web mining. A three-layered framework for ontology driven web mining has been proposed. The functionality of the layers in the conceptual framework is also explained. The traditional clustering method is unified with ontology in the proposed layered architecture[23].

- To study the role and advantages of semantics and ontology in effective web content mining and information retrieval. The concept and role of ontology in mining has been discussed. The advantage of ontology inclusion in clustering process has been described. [24,25]

- A comparative analysis of clustering method with and without WordNet ontology has been performed to evaluate the result. The implementation of clustering method has been done in R language.
1.4 Outline of the thesis

Chapter 1 Introduction

The chapter presents introduction to web mining, clustering and ontology. The limitations of keywords based retrieval are discussed. The motivation and objective of the work has been explained. The role of ontology in web mining has been briefly described. The chapter also lays emphasis on the proposed layered architecture for web mining. The advantage and improvement of including WordNet ontology in clustering methods is discussed in the chapter.

Chapter 2 Background and Related Work

Chapter 3 Finding Clusters of Data: Cluster analysis in R

Chapter 4 Natural Language Processing techniques applied in Information Retrieval- Analysis and Implementation in Python

Chapter 5 Ontology driven Web Mining: An Enhanced Approach of Preprocessing the Document using WordNet in Text Clustering

Chapter 6 Conclusion and Future Directions

Chapter 7 Publications and Papers Presented

It contains the papers that have been written and published/presented in the international/national conferences and reputed journals while carrying out the proposed work.