Due to the huge volume of digital data having different length and nature, graph based intelligent text mining systems have become a great utility for knowledge discovery. An important aspect of the graph-based method is that it does not require deep linguistic knowledge and it does not depend on domain or language specific annotated corpora. These features make it highly portable to other domains, genres, or languages (Mihalcea, R. 2005). Additionally, graph based techniques works well because they do not only rely on the local context of a text unit (vertex) (Aggarwal and Zhai, 2012), but they also take into account information from the entire text (for example: (1) In graph models, such as LexRank, the entire document is represented as a network of interrelated sentences (Erkan and Radev, 2004), (2) words, phrases named entities and sentences, etc., can be used as vertices to represent text (text collection) as a graph (Nenkova and McKeown, 2012), (Aggarwal and Zhai, 2012). We can easily apply probabilistic models and linguistic features at the top of the graph based model to enhance the quality of the extracted information for linguistically reach documents (Aggarwal and Zhai, 2012).

Graph based techniques have been effectively applied in several areas of text mining. According to Nenkova and McKeown(2012), Graph-based approaches have been shown to work well for both single document and multi-document summarization. Since, the approach does not require language-specific linguistic processing beyond identifying sentence and word boundaries; it can also be applied to other languages, for example, Brazilian Portuguese, etc. At the same time, incorporating syntactic and semantic role information in the building of the text graph lead to superior results over plain TF*IDF induced cosine similarity (Chali and Joty, 2008).

A lot of time text data supports a network of information. For example: (1) Email messages are linked to the senders and recipients, (2) Publications are also linked by citations (3) Many documents are related to a social network and so on. The analysis of these documents can reveal more interesting structure if the network graph can be incorporated (Aggarwal and Zhai, 2012). Shen and Lapata(Shen and Lapata, 2007) assess the contribution of automatic semantic role labeling (Gildea and Jurafsky, 2002) to factoid QA, by treating semantic role assignment as a global optimization problem in a weighted bipartite graph, and answer extraction as an instance of the graph matching problem. Mihalcea and Tarau(2005), describes a method for language independent extractive summarization that relies on iterative graph-based ranking.
algorithms. Mihalcea and Tarau (2004), uses word graph and page-rank based scheme for keyphrase extraction also known as “TextRank”. It was further extended towards automatic text document summarization. Now, "TextRank" is considered as one of the state-of-the-art in keyphrase extraction.

However, still some unexplored areas are left. Next, a lot of explored areas require sincere attention towards improvements. In this thesis, we focus on some core issues of text mining (and techniques for them) to solve these problems. For this, we have identified some core text mining issues like: (i) meaningful phrase identification (ii) differentiating role and sense of words, preferably via a single measure, (iii) handling information gap at the phrase level by using unsupervised scheme, (iv) integrating the importance of words as a core feature and (v) identifying group semantics and/or logically related features, (vi) sentence abstraction and so on. We consider it as an intelligent text mining, i.e., the discovery or creation of new knowledge from a collection of documents. The new knowledge may be the statistical discovery of new patterns in known data (standard text mining) or it may incorporate artificial intelligence abilities to interpret the patterns and provide more advanced abilities (intelligent text mining, (Kroeze et al. 2003)).

It is important to note that, all languages having basic computation facilities contain such minimum linguistic information (i.e., stop-words, stemmers and punctuation marks). This gives strong background support to our way of using limited linguistic dependencies. This feature enhances the applicability/scope of such systems to work with both types of languages, i.e., languages which are computationally-reach and languages which are computationally poor (having minimum computation facilities). Another important issue is the effectiveness of such techniques over supervised, linguistic and domain information supported techniques, which we achieved.

Chapter organization: In section, 1.1, we present the core issues that we have focused on the thesis. In Section 1.2, we present our proposed text mining framework. The last section (i.e., section 1.3) contains brief description about thesis organization.

1.1. Core Issues or Problem Addressed in the Thesis

In this section, we cover the following core issues, which have remarkable impact on several text-mining applications. Some of them, e.g., (i) meaningful phrase identification (ii) handling information gap at the phrase level by using unsupervised scheme, (iii) integrating the importance of words as a core feature and (vi) sentence abstraction requires sincere attention to the improvements in the performance. While, some
of them, e.g., (i) differentiating role and sense of words, preferably via a single measure, (ii) handling information gap at the phrase level by using unsupervised scheme, and (iii) identifying group semantics and/or logically related features, are newly explored issues. Finally, (i) handling information gap at the phrase level by using unsupervised scheme, and (ii) sentence abstraction are little explored and require remarkable improvements. The following contains the brief discussion about application specific requirements of the discussed core issues and our contributions to solve the research problems of the discussed core issues.

1.1.1 Identification of Meaningful Phrases/ key N-grams

Identification of sophisticated/meaningful phrases (Key N-grams) is very important for several text mining applications. This is typically because the local context of the words remains preserved in the phrases and this context is usually beneficial if the semantics of the word is dependent on its locality. For example, “Taj Mahal hotel” and “Taj Mahal tea” do not have anything to do with the Taj Mahal (the seventh wonder of the world). Apart from the use of meaningful phrases or key N-grams as a candidate keyphrases in keyphrase extraction, it can be useful in several other applications.

According to (Wang et al. 2007), word order and phrases are often critical to capturing the meaning of text in many text mining tasks and hence concentrated on the development of topical N-grams and topical phrases for improvement in the performance of information retrieval processes. Similarly, (Wang et al. 2012) used search focused key N-grams for relevance ranking in web search. (Bergsma et al 2009), used N-gram based model for lexical disambiguation. Additionally, the successful use of N-gram based model in Machine translation (Gong et al. 2012), and automatic text classification (Bergsma et al 2009), etc., also motivate us to concentrate on the identification of meaningful phrases or key N-grams from documents. Now, we will go through issues regarding the qualities of the identified phrases, language dependencies, approach applied and related applications.

Issues regarding Quality of Identified Phrases: Most of the traditional phrase identification system suffers the problem of (1) partially correct phrases, i.e., phrases with missing starting or end word(s) or (2) phrases with additional noisy words at starting or end position. (3) The direct use of N-grams suffers the problem of multiple phrases. (4) Using parts-of-speech and regular expression also results in several phrases. (5) Another, but very important issue about the phrase identification scheme is: “What may be the optimal number of phrases, required to effectively represent the information content of document”. Solving these problems can improve the quality of different text mining applications.
**Language Dependencies:** Language dependency is a strong barrier, which prevents us in extending the applied scheme for multi-lingual environments. Thus, reducing the language dependency in phrase identification became an interesting task of significant importance.

**Brief Approach:** To solve this problem, we concentrate on a combined approach which uses statistical and semantic presence of noisy or less informative words around candidate key N-grams. Due the relatively high occurrence frequency and/or different use of noisy or less important words around the candidate key N-grams, it becomes easy to capture the key N-grams from the given word sequence. To achieve this goal, we use a word graph representation of documents (Kumar et al. 2015). Next, to identify the phrase boundary, we use betweenness centrality and the page rank score of nodes of the graph. The proposed technique is very helpful in keyphrase extraction and document clustering (where, we used phrase based document vector modal to cluster the documents).

**1.1.2 Differentiating Role and Sense of Words**

According to Abdalgader, K. (2014), word sense identification improves the measurement of short-text similarity. Word Net and Babel Net, etc., contain manually compiled thesaurus which lists the different sense of each word, as well as word relationships, such as synonymy, and antonyms etc. However, the limited size is a bottleneck of such systems.

Identifying “role” and “sense” of the same words in two or more documents are challenging task. The importance of such task increases, when given document collection contains a large number of matching words. As, in several cases words may have the same sense in two different sentences or documents, but having a different role.

**Approach and applications:** To solve this problem, we propose a graph based mapping of co-occurring words and closeness centrality score (Kumar et al. 2012a). This technique utilizes global information w.r.t. N-gram or phrase pattern based matching or probabilistic measures etc. We use this technique in automatic evaluation of machine generated summaries, guided answers and essay evaluation, where “role” and “sense” of words play a very important role. As, a future work, such technique can be applied in document clustering.

**1.1.3 Reducing Information-Gap between Phrases**

Several times, topically similar documents contain conceptually similar but non-matching phrases. This may be due to differences in writing scheme or strategies (e.g., use of the phrase “AK-47 assault rifle” in
one document and “Kalashnikov rifle” in the other document). This is an important problem and badly affects the performances of most of the text mining algorithms.

To solve this problem we, calculate Wikipedia anchor text community and apply it to identify the conceptually related terms. We use this concept in document clustering (Kumar et al. 2010).

1.1.4 Exploiting Importance of Words

Most of the data mining techniques treat word or phrases as data points and thus ignore their importance. For example:

<table>
<thead>
<tr>
<th>Table-1.1: A sample text (to show the importance of words)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample text on “Social Networking”; source Wikipedia:</strong></td>
</tr>
<tr>
<td>A social network is a social structure made up of a set of actors (such as individuals or organizations) and the dyadic ties between these actors (such as relationships, connections, or interactions). A social network perspective is employed to model the structure of a social group, how this structure influences other variables, or how structures change over time.[1] The study of these structures, uses methods in social network analysis to identify influential nodes, local and global structures, and network dynamics. Social networks are distinct from information, biological, or electrical networks, but theories and methods generalizing to all of these complex networks are studied in the field of network science.</td>
</tr>
</tbody>
</table>

In this sample text (see Table-1.1), we cannot equally treat every word. Even some bold face words seem to have more importance than other normal words. If we carefully read the text, then we find different levels of importance of different bold face words.

Thus treating every word or word sequence as a separate data point may be inappropriate and known data mining techniques may not be applied straightway. Actually, keyphrase extraction techniques use several statistical, linguistic, grammatical facts and heuristics to score the words according their level of importance.

**Approach and Applications:** We use, the importance of words in the calculation of (1) similarity measures for document clustering (Kumar et al. 2010), (2) in the calculation of weighted clustering coefficient for document summarization (Kumar et al. 2012b) and (3) in the calculation of importance of topics in summarization, summarization evaluation (Kumar et al. 2012a) and evaluation of descriptive answers etc., (Kumar et al. 2010a).
1.1.5 Identifying Group Semantics or Logical Relatedness

Actually, we introduce the term group semantics/ logical relatedness to represent the semantically similar or useful terms w.r.t., a set of terms. Here each term in the given set of terms may have different semantics and may not be related to each other directly. For example, suppose we want all terms which logically define the blue color of the sky. The answer may be terms like “air scattering”, “short-wavelength”, “light” etc. It is important to note that, these terms are not directly or semantically related to either “sky” or “blue” but logically related to the term “blue color of the sky”.

We introduce such terms as, group semantics/logical relatedness, i.e. they are related to the concept carried by the set of terms.

**Approach and applications:** To identify the terms which are semantically related to the entire group, we have introduce the use of semantic correlation graph and modified version of personalized page rank using random walk with restarts. The proposed technique is very effective for “WHY based question answering” and can be further explored for some other applications (Kumar et al. 2013a).

1.1.6 Sentence Abstraction

Actually, sentence abstraction is one of the least explored, but very useful techniques for document summarization. When humans produce summaries of documents, they do not simply extract sentences and concatenate them. Rather, they create new sentences that are grammatical, that cohere with one another, and that capture the most salient pieces of information in the original document (Knight and Marcu, 2002)). They devised both a noisy-channel and a decision-tree approach to solve the problem. Similarly, Filippova, K. (2010), consider the task of summarizing a cluster of related sentences with a short sentence which she call multi-sentence compression and present a simple approach based on shortest paths in word graphs. However, still (1) grammaticallity and (2) the presence of important pieces of information is a challenge for all such sentence abstraction related tasks.

1.2 Proposed Text Mining Framework

The proposed text mining framework uses core text mining techniques for various applications. We propose total five different core text mining techniques (see sub-sections 1.3.1 to 1.3.5). The proposed techniques use graph representation of document(s) and limited linguistic resources (i.e., limited to, include only stopwords, stemmers and punctuation marks). To properly handle the requirements of
different types of applications, we enhance the traditional word graph based representation (Mihalcea and Tarau, 2004) of text by populating it with appropriate combination of statistical, semantic and topical information. We also introduce the phrase (N-gram, n>=1) graph based representation of text. In phrase graph representation, we treat every distinct phrase as node of the graph.

From figure 1.1, it is clear that core technique/issues, like: “Integrating importance of words with text mining techniques” is applicable to all five applications (i.e., automatic document summarization, automatic summarization evaluation, document clustering, keyphrase extraction and automatic question answering), technique like: “Differentiating role and sense of the words” is applicable to automatic summarization evaluation, but the chances of their use in other application like document clustering is very good. Similarly, techniques, having few related applications at present, have a good chance of wide applicability. This all proves our view of concentrating on development of core text mining techniques for multiple text mining applications.

1.3 Our Contribution

As discussed in previous sections, we propose intelligent text mining techniques, which uses limited language dependencies (i.e., limited to, include only stopwords, stemmers and punctuation marks) and avoid the use of domain information. Our proposed model uses simple input cleaning and preprocessing task. All the core techniques developed by us, use word graph based representation or phrase graph based.
representation of text. We also use the Wikipedia knowledge resources to enhance the quality and usability of document representation.

Finally, we introduce the use of social networking/graph based techniques with a lot of additional features to reduce the language specific dependencies and to enhance the efficiency of result in text mining. We have also introduced/improved some supporting techniques (graph based) for basic text mining applications. Thus, we can divide our contribution into the following two categories:

**Major contribution:**

1. **Meaningful phrase identification:** We introduce the graph based N-gram (N\(\geq 1\)) filtration technique, which uses the combination of statistical and semantic features for meaningful phrase identification. We use it in Keyphrase extraction and document clustering.

2. **Differentiating role and sense of words, preferably via a single measure.** We introduce the graph based common measure to identify the role and the sense of the words. We use it in automatic evaluation of machine generated summary.

3. **Handling information gap at the phrase level by using unsupervised scheme.** We introduce the use of Wikipedia anchor text structure and community detection scheme, etc., to handle the information gap at the phrase level (which occurs due to the difference in writing strategies). We use it in document clustering.

4. **Integrating the importance of words as a core feature.** Instead of applying equal treatment to words, having different levels of importance, we introduce the use of weighted importance of words in the text mining applications. We use it in the entire thesis. However, the calculation of the importance of words differs and depends upon the application specific requirements.

5. **Identifying group semantics:** The main focus of this technique is to identify the logically and/or semantically related words for the given set of words. We introduce the use of improved version of personalize page rank using random walk with restarts on the semantic correlation word graph of text (populated with statistical and semantic information) to calculate the group semantics for the given set of words. We use it in Why-based question answering system.

6. **Sentence abstraction:** We introduce a vertex constrained shortest path based sentence abstraction technique, which combines the important information spread in different sentences into a single sentence and thus increases the informativeness. We use it in document summarization (single and multi-document summarization).
Additional/Supporting Contributions:

1. We introduce a fast method for the extraction of domain specific/nearest-neighbor/ contextually similar DBpedia extended abstracts for any given text or text collection. We use the domain specific/nearest-neighbor/ contextually similar DBpedia extended abstracts for calculation of semantic relatedness score for word pairs of the given text. The effective combination of offline and online component of the proposed technique has made the system faster. The use of domain specific/nearest-neighbor/ contextually similar DBpedia text in the calculation of semantic relatedness has some additional benefits. As, it is free from huge variations in the use of terms in Wikipedia like huge knowledge resource.

2. To capture the information coverage/reachability of keyphrases in semantically reach N-gram (N>=1) graph and to reduce the effect of noisy terms, we introduce the use of the information theoretic model for candidate key player detection (Oritz-Arroyo, D., 2010). The proposed approach uses the combination of entropy based measures and closeness centrality score of keyphrases.

3. Based on the requirements of the different text mining applications, and proposed core techniques, we have investigated and explored different types of graph formation for text documents. For example the use of (1) undirected word graph, (2) forward directed word graph, (3) reverse directed word graph and (4) N-gram (N>=1) graph. In most of the cases, we populate the graphs with statistical and semantic information.

4. We introduce the use of weighted minimum vertex cover based technique for extraction of topical terms from text for single document summarization.

Scope of our proposed approach

Our proposed graph theoretical model can be served as a basic model towards the development of several intelligent text mining techniques.

Due to the use of graph based techniques, our proposed model has a lot of technical advantages like (including but not limited to): (1) This model provides a very simple representation of tokenizable text, (2) we can take advantage of research-rich graph theoretical background to enhance the efficiency of the system and (3) we can easily put the graph theoretical model on the top of other models like: probabilistic model, statistical model and linguistic model etc.,. These features improve the flexibility of such techniques.
Similarly, due to the limited dependencies on linguistic resources, the proposed model can be easily extended to other languages and domain environments.

1.4 Thesis Organization

As we discussed earlier, the contributions of this dissertation are the development of domain independent and less language dependent core techniques. We applied these techniques in text mining applications like: (1) Keyphrase Extraction, (2) Document summarization (Single and Multi document summarization), (3) Automatic Summarization and Descriptive Answer Evaluation, (4) Document Clustering and (5) Automatic question answering. However, we present the entire work in the form of independent chapters, where each of them is dedicated to separate text mining application (see above). Each of these chapters uses some enhanced graph theoretical techniques, which are based upon the core techniques, discussed above. The chapters are organized as follows:

Chapter 2: In this chapter, we discuss the basic text-mining framework and present a brief literature survey to cover the novel research efforts in different areas of basic text mining. We also explore the important and crucial research issues with basic text mining techniques. Finally, we present some novel and effective research efforts and their pros and cons.

Chapter 3: In this chapter, we introduce the use of weighted minimum vertex cover based technique for extraction of topical and thematic sentences for single document extract summarization. This scheme uses the weighted importance of words in document. We also present single document abstract summarization system, which uses vertex constrained shortest path based scheme.

Chapter 4: presents unsupervised approach for multi document summarization. In this chapter we propose the use of weighted clustering coefficient, page rank and task focused use of Wikipedia for multi-document summarization. We also extended the sentence abstraction scheme used in the single document summarization system to multi-document summarization.

Chapter 5: presents techniques for Automatic Summarization Evaluation. In this chapter, we focus our attention towards the use of (1) importance of words, (2) importance of topics, (3) role and (4) sense of words in evaluation of machine generated summary, guided answers and essays.

Chapter 6: presents the techniques for unsupervised keyphrase extraction. This chapter focuses on two different tasks, i.e. (1) identification of candidate N-grams (n>=1)/phrases and (2) ranking of the identified phrases.
**Chapter 7**: presents document clustering application. Here we use (1) importance of words in documents in similarity measure and (2) Wikipedia anchor texts for improvement in phrase correctness and reduction of information gap between phrases in different documents. As, discussed earlier, due to difference in writing strategies, we may use different phrases for similar topic in different document, which may result in the information gap. The presence of such information gap reduces the performance of document clustering.

**Chapter 8**: In this chapter, we use the technique of group semantics/logical relatedness to develop the application, which can automatically answer “Why” based questions. However, we also used the importance of words to reduce the chances of getting higher importance by noisy or less important terms.

Finally, in **Chapter 9**, we present the conclusion of the entire work.