Chapter 8

Conclusion and Future Work

In today’s digital era, most of the textual information is produced by academicians and research community through international conferences, journals, workshops and other publications covering every field of natural sciences and technology. This is expanding already huge repository of scientific knowledge at an alarming rate. The new researchers are exposed to this huge amount of information available in the form of published research papers to get latest research trends and developments in all relevant subject areas. But the information overload problem makes it difficult to assimilate even a portion of the potentially related material. Moreover multiple domain specific research papers from different publications mostly have similarity and repeatability in contents. Thus reading all research papers from end to end to get the latest developments in the interested domain becomes time-consuming, unnecessary, irrelevant, cumbersome and sometimes impossible. Here we felt that a tool for mining most relevant contents from domain specific papers was badly needed.

Basically the reader focuses in the previously published papers ‘main research problem’ and ‘the approaches’ used to solve these problems. When a person tries to initiate research in a new field he/she tries to understand the current state of research going on, standard and general approaches described, relation to other’s work, overview of the evaluation methodology, typical numerical results in the field, its advantages, drawbacks, conclusions and future scopes. At this stage the new researcher or even experienced researcher cannot be sure whether retrieved papers are relevant to him or not. On the other hand, even more frustrating thing is to read each and every paper to search for the necessary information. However, reading through irrelevant material more or less randomly is a cumbersome process wasting reader’s valuable time.
After realizing that the whole exercise was tiresome, we have tried to provide an automated summarization tool for solving above problems. This innovative ‘Optimized Summarization System’, (OSS) successfully provides search relevant categorical summary of single as well as multiple research papers relevant to a specific domain reducing redundancies.

We have observed that the techniques for automatic shortening of texts such as extracting salient sentences from text using features like word and phrase frequency [3], position in the text [34], key phrases [4] and text processing [38] have concentrated mostly on newswire data [39, 40] only. Many approaches addressed the problem by building systems depending on the type of the required summary. But these solutions are not useful for general purpose summarization. Even the abstractive summarization using advanced language generation techniques [41, 42, 11], decision tree learning (C4.5) and the Bayesian classifiers [9, 10] cannot handle heterogeneous sentence features effectively when such features are unknown.

Previous approach of concept identification and presentation [104] concentrated on linguistic expressions like nouns, verbs and adjectives, summarizing multiple articles for generic use. But this approach did not consider redundancy and constructing such expressions became time consuming. OSS addressed redundancy problem using ‘Maximal Marginal Relevance’ (MMR) and ‘Research Relevant Novel’ (RRN) term identification criteria.

Existing summarization systems for various textual based applications use different approaches suggested by many authors. However they do not address the issue of summarizing research articles. Various learning and semantic-based methods were also used in DUC’s task of creating N-words summary for news articles. But the best performing systems [13] were not able to reach the standard baselines set by DUC. This shows that even the methods suggested were not able to solve the given task.

Many authors suggested summarization methods for relevance judgment of documents [68, 69, 70]. Some of the suggestions were Query Biased Summaries, Summaries Based on Lexical Chains using Information Retrieval Tasks, Task-based text summarization and customized summaries as per requirements. But these approaches do not attempt to analyze the source in great detail; instead they used training to extract snippets containing important information.
We studied and addressed all the above summary generation approaches, issues and came up with the idea of using Data mining techniques for solving a distinct task of domain specific research papers summarization. We used various data mining techniques such as preprocessing which separates the ‘Abstract’ and ‘Introduction’ sections from paper, identifies boundary of each sentence and then tokenizes them. The term identification criteria will then search for all the research relevant novel terms which are present in the sentences. The clustering technique clubs extracted sentences under five RRN categories such as Research Goals, Research Methods, Research similarity/dissimilarity, Research Continuation/Novel work and Research Outcome. The sentence scoring method compares RRN terms present in the sentences and then ranks them. Finally the research relevant sentences are presented in summarized form under five categories as per user need by controlling the summary length.

After comparing with earlier summarization approaches our system results conclude that,

- The collection of innovative RRN terms helps in identifying the role of each research oriented sentences under various user expected search categories. The search categories like research goals, research methods/approaches, continuation of existing research or novel system, similarity or dissimilarity with other systems and research outcomes are most user centric. This completely differentiates OSS from existing user centric summarization approaches.

- The sentence extraction data mining approach under various research categories for automatically retrieving most related sentences helps in identifying important material from text which has a significant and complete relative meaning.

- This technique replaces conventional iterative extraction methods, which extract one sentence at a time without considering the rest of the sentences in the summary, with more holistic ones, where the decision to extract a sentence is determined not only by the content of a single sentence, but also by the rest of the extracted sentences.

- The use of clustering in a data mining process automatically groups most related multiple documents into clusters or bunches. Thus the small collections containing only related
categories are searched, improving effectiveness and efficiency. Clustering give users an overview of the contents of a document collection reducing the search space.

- The use of optimization process enhances effectiveness of automatic summary, to make it function at its best. The optimized summary presents innovative contribution through most useful sentences from papers. It significantly improves the conciseness of automatic summaries by cutting unwanted, repeated, non-related material and presenting only the abstract.

- OSS saves reader’s time by avoiding unnecessary material reading and understanding for which he is not interested. This process also shortens system output and minimizes space requirements.

- Thus OSS achieves remarkable progress avoiding system redundancy and providing the reader with related research material only.

This innovative OSS tool developed using Data Mining strategies, is an aid for research scholars for getting short, condensed, accurate, explicit, optimized and most relevant summarized information from domain specific topic-oriented multiple published research papers. This research also highlights earlier and latest research developments, progress, challenges and future scope in the particular field of study. It provides initial material for further innovation specifying research methods/techniques/approaches used and comparison with others. Here sentence clustering strategy automatically determines the number of clusters where the summary length is selected by user. The sentence selection method consisting RRN terms is presented and compared with other local methods.

Few advantages of OSS as compared with other existing approaches are,

- Research papers Abstract and Introduction section summarization is sufficient for the reader to decide whether he/she should read the complete paper to move ahead or not.
- OSS evaluations show steady correlation with the human assessment outcome.
- Descending order sentence score comparisons give the most optimized results.
- Optimization enhances the effectiveness of OSS to make it function at its best.
• Reduction significantly improves the conciseness of OSS.
• OSS introduces new information criteria by the use of dividing document into research categories thus providing new structure for summarization which improved output readability as compared with other systems.
• OSS has high interpretability and extensibility.
• Exploring this automated system comes at a timely moment, as scholars now have access to huge amount of scholarly literature.

From the results it is clear that there is no such thing as the perfect summary for all purposes. Rather, the quality of a summary is subjective, task dependent, and possibly specific to a user. Use of RRN term analysis and query term show that the term-based summarizer outperforms the word-based, which confirms that term or feature, is more than the word.

**OSS Limitations:**

The downside of our approach is the fact that the OSS system is domain dependent, which means it cannot be easily ported to different domains or to be used in domain independent contexts. Thus designing the domain independent summary using the same approach for genre use can be our future task. Furthermore, since summary length is largely determined by thread, system output must be further processed to generate a summary of a desired length. Also the output of our system consists of purely extracted sentences, where we do not perform any sentence compression or sentence generation.

However, the coherent text is only identified by Rhetorical Structure Theory [36]; this can help extracting two or more discourse sentences. Since the summary contains only some of these sentences resulting incoherency.

Recently, the concept of computer-aided summarization was developed as a means of integrating the human and automatic summarization processes [107]. This approach presents a feasible alternative for users, recognizing the shortcomings of fully automated process and allowing the human summarizer to interact with it producing the best possible summary.

More recently few authors argue that citation texts are useful in creating a summary of the important contributions in a research paper [61]. The citation text of a target paper is the set of
sentences in other technical papers that explicitly refer to it; however using citation text directly is not suitable [105].

**Future Scope:**

OSS needs further modifications to overcome some of its weaknesses which will be the improvements to be carried out to function it as one of the best summarization tool. Few limitations of OSS are as follows,

- As mentioned earlier, a wide range of different sub-domains of the specific field can be covered by adding few modifications into this system. With the five categories described in this theory, it is not designed to model the full complexity of all research articles.
- These categories can be further subdivided into many for simplicity. The system is unable to identify the tag lines given between Abstract and Introduction sections.
- All sections of the research papers can be covered to produce in depth summary. This needs standardized formatting of topic and sub-topic headings.
- We have not considered citations [106]. Inclusion and understanding the cited text may improve summary contents.
- In order to benefit fully from such a system, users (who are not necessarily experts in summarization or linguistics) need to know exactly how best to improve the automatic output to transform it into a high quality, coherent and readable summary.
- A summary can have a completely different meaning and structure (e.g. tables and graphs are used in economics, a map summarizes the main features of a country, whereas a trailer is a summary of a film or TV show), but detailing them is beyond the scope of this thesis.
- New applications for summarization, such as question-answering, condensation and navigation of book-length materials, summaries for hand-held devices, etc., will create new opportunities as well as challenges for summarization.