Chapter 4

System Description

Overview

This chapter provides the innovative contribution of this thesis. The basic study is in ‘Research Related Novel Terms’ identification and categorization. The term based summarization system architecture is presented using different data mining strategies such as sentence extraction, clustering and optimization. The application areas, the main building blocks and detailed system implementations are described.

4.1 System

This new system is named as ‘Optimized Summarization System’ (OSS). OSS is based on the human and term-based summarizer. This approach can easily be integrated with other methods to get the better performance. Using this approach the system first identifies the structure of the source. In order to achieve this, the source’s title, outline, layout and table of contents are examined. This system minimizes the information overload problem by providing a hopeful solution as automatic reduction, optimization and then summarization of domain specific topic-oriented multiple related research papers retrieved through search engines according to scholar query.

4.1.1 Innovative Contribution of the Thesis

The main original contributions of this thesis fall into three categories:
1. ‘Research Related Novel Term’ (RRN) identification and categorization
2. Use of MMR [67] criteria and
3. Data Mining Strategies for automatic optimization and summary generation from multiple related research papers according to scholar expectations.

This innovation identifies RRN terms. The RRN terms are the words or phrases indicating research contribution of the paper elaborating research categories and having relative complete meaning. The reasons behind developing summarization system around research oriented terms are,

- Terms are one of the most common components of automatic summarization systems developed till date; therefore, a clear presentation of terms produces an impact in the field under study.
- Term-based summarizer can be quickly implemented and relatively easily integrated with other already existing or developed methods.
- As a result, it is possible to experiment with all multiple research oriented term as parameters.
- Finally term based summarization strategies makes this system application oriented.

In this work we aimed to test the RRN term analysis based approach on a set of related research papers by identifying innovative contribution in the paper guiding research scholar. The RRN terms that provided basis for the summarization procedure, have been constructed around the research paper’s novelty and identified words that highly correlated with it. To form a domain specific topic oriented summary, this system selects those sentences that contained the research related novel words. By selecting only research related categorized information this system fulfill many searchers' information needs, particularly the needs of research scholars.

4.1.2 Research Categories

The research papers are made available by many digital library systems or search engines as electronic text documents on account of a scholar query. These multiple papers are most related having similar and repeated contents and do not provide condensed, accurate and most topic
related information requested by the scholar. Mainly all research papers contains essential information about,

1. Innovative scientific ideas (research aims/goals),
2. Uniqueness and difference from previous ideas (research similarity/dissimilarity),
3. Approaches & methodologies (research methods/approaches),
4. Research continuation of earlier/existing work (research continuation/novel) and
5. Results & discussions (research outcomes).

The most important and interesting point in the list of related research papers is the domain specific topic oriented innovative contributions. Thus OSS automatically summarizes each paper’s as well as multiple paper’s research purposes/aims, approaches & methodologies used, existing research continuations or novel work, authors own and other researcher’s work and results & discussions. Therefore by identifying the number of contributions into the interested area helps a scholar to find the progress, challenges and future scope in that particular field of study for quickly moving into the new research area.

To identify individual research contributing statements from single and multiple papers, informing research aim, similarities or contrast, research methods, research continuation and research outcome only ‘Abstract’ and ‘Introduction’ sections of the whole documents are considered. The reason behind selecting only these two sections is that all the research contributing statements are mostly covered in these two sections completely. Other sections contain the same statements repeatedly with some more or less words including conclusion section also. Thus summarizing only ‘Abstract’ and ‘Introduction’ sections from lengthy research papers give advantage of reduction and optimization of scholar efforts, time and choice of relevance.

4.2 Sentence Role

A summary helps us to judge the relevance of a document. Many researchers have studied summarization methods for relevance judgment of documents [68, 69, 70]. There are many different algorithms for finding related documents. We use the algorithm for relevance by the use of RRN categories and the sentence roles. This algorithm focuses not only on the run time as
commonly done in on-line search engines, but also on finding highly related document categories.

By studying many research papers, we identified that all the basic research oriented statements are almost covered in ‘Abstract’ and ‘Introduction’ sections dealing with the scholar search. All the research innovations are stated by multiple sentences starting with or including most of the RRN terms. These terms can be categorized into various sentence roles. The overview of these sentence role categories shown by ‘Abstract’ and ‘Introduction’ sections concentrating on ‘RRN’ are shown in figure 4.1.

<table>
<thead>
<tr>
<th>Research Category</th>
<th>Sentence Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research Goals</td>
<td>Sentences representing the purpose or aim or principle innovative idea of research under study for current paper;</td>
</tr>
<tr>
<td>Research Methods</td>
<td>Sentences representing the methods or approaches or ways used for the goal achievement;</td>
</tr>
<tr>
<td>Research Similarity/Dissimilarity</td>
<td>Sentences claiming authors own work contrast with others/ earlier work;</td>
</tr>
<tr>
<td></td>
<td>sentences showing limitations in others/ earlier work;</td>
</tr>
<tr>
<td></td>
<td>direct comparing with others/ earlier work;</td>
</tr>
<tr>
<td></td>
<td>research work of this kind never done before;</td>
</tr>
<tr>
<td></td>
<td>sentences presenting similarity with others / earlier work;</td>
</tr>
<tr>
<td>Research Continuation/Novel</td>
<td>Sentences describing research continuation of earlier/existing work;</td>
</tr>
<tr>
<td>Research Outcome</td>
<td>Sentences relating to result, conclusion, outcome;</td>
</tr>
<tr>
<td></td>
<td>Sentences showing end product;</td>
</tr>
<tr>
<td></td>
<td>Sentences stating evaluation of implementation;</td>
</tr>
</tbody>
</table>

Figure 4.1. Overview of sentence role describing ‘RRN’ terms.

4.3 Research Related Novel Terms

‘Research Related Novel Terms’ (RRN) are identified manually by studying many published research papers and their content management structure. Identified RRN terms are then
categorized into five research categories according to their sentence roles. Manually identified RRN terms under research categories are shown in figure 4.2.

<table>
<thead>
<tr>
<th>Research Category</th>
<th>Identified RRN terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Research Goals</strong></td>
<td>the aim of this</td>
</tr>
<tr>
<td></td>
<td>the purpose of</td>
</tr>
<tr>
<td></td>
<td>we observe</td>
</tr>
<tr>
<td></td>
<td>argue that the</td>
</tr>
<tr>
<td></td>
<td>this paper</td>
</tr>
<tr>
<td></td>
<td>paper explore</td>
</tr>
<tr>
<td></td>
<td>study the problem of</td>
</tr>
<tr>
<td></td>
<td>paper provide</td>
</tr>
<tr>
<td></td>
<td>research goal</td>
</tr>
<tr>
<td></td>
<td>the problem addresses</td>
</tr>
<tr>
<td></td>
<td>address the issue of</td>
</tr>
<tr>
<td></td>
<td>address the problem</td>
</tr>
<tr>
<td></td>
<td>Design</td>
</tr>
<tr>
<td></td>
<td>literally</td>
</tr>
<tr>
<td></td>
<td>explore the use of</td>
</tr>
<tr>
<td></td>
<td>defines as</td>
</tr>
<tr>
<td></td>
<td>to make up</td>
</tr>
<tr>
<td></td>
<td>give an overview of</td>
</tr>
<tr>
<td><strong>Research Methods</strong></td>
<td>mechanics suggests</td>
</tr>
<tr>
<td></td>
<td>algorithm</td>
</tr>
<tr>
<td></td>
<td>new method</td>
</tr>
<tr>
<td></td>
<td>architecture</td>
</tr>
<tr>
<td></td>
<td>system consist of a</td>
</tr>
<tr>
<td></td>
<td>a powerful tool for</td>
</tr>
<tr>
<td></td>
<td>our experiments</td>
</tr>
<tr>
<td></td>
<td>method</td>
</tr>
<tr>
<td></td>
<td>we propose</td>
</tr>
<tr>
<td></td>
<td>Scheme</td>
</tr>
<tr>
<td></td>
<td>in the field of</td>
</tr>
<tr>
<td></td>
<td>new way</td>
</tr>
<tr>
<td><strong>Research Continuation or Novel</strong></td>
<td>contrast to</td>
</tr>
<tr>
<td></td>
<td>in comparison with</td>
</tr>
<tr>
<td></td>
<td>unlike</td>
</tr>
<tr>
<td></td>
<td>in the context of</td>
</tr>
<tr>
<td></td>
<td>work is based on</td>
</tr>
<tr>
<td></td>
<td>the existence of</td>
</tr>
<tr>
<td></td>
<td>does not depend on</td>
</tr>
<tr>
<td></td>
<td>History</td>
</tr>
<tr>
<td></td>
<td>Existing work</td>
</tr>
<tr>
<td></td>
<td>as stated by</td>
</tr>
<tr>
<td></td>
<td>system like</td>
</tr>
<tr>
<td><strong>Research Continuation or Novel</strong></td>
<td>Novel</td>
</tr>
<tr>
<td></td>
<td>Novel</td>
</tr>
<tr>
<td></td>
<td>in addition to work</td>
</tr>
<tr>
<td></td>
<td>seems to agree with</td>
</tr>
<tr>
<td></td>
<td>approach carried over</td>
</tr>
<tr>
<td></td>
<td>in previous work</td>
</tr>
<tr>
<td></td>
<td>can be extend to</td>
</tr>
<tr>
<td></td>
<td>research extension</td>
</tr>
<tr>
<td></td>
<td>in our first experiment</td>
</tr>
<tr>
<td></td>
<td>refers to</td>
</tr>
<tr>
<td></td>
<td>we use the framework</td>
</tr>
<tr>
<td><strong>Research Outcome</strong></td>
<td>prove that</td>
</tr>
<tr>
<td></td>
<td>Performance</td>
</tr>
<tr>
<td></td>
<td>can be used to</td>
</tr>
<tr>
<td></td>
<td>establishing research</td>
</tr>
</tbody>
</table>

SVKM’s MPSTME, NMIMS, Mumbai
Output | field | demonstrate that
---|---|---
at the end we state | investigation | end product
general claim | contribution | Result
in the form of | outcome | carried out
Determine | study show that | the evaluation
the result show that | result indicate | Implementation
believe that | conclude | Conclusion

**Figure 4.2.** Research Related Novel Terms (RRN) for research categories.

The ‘Abstract’ section of research paper focuses on:

i. The purpose of research; where-in the statement of the problem(s) or research issue(s) are addressed with the research methods/approaches/techniques used (experimental research, case studies, questionnaires, etc.);

ii. The results and/or findings of the research; and

iii. The main conclusions and recommendations, therefore, briefly summarize the work done.

And the next ‘Introduction’ section provides:

i. General description about the importance of the topic and its history in the field;

ii. Establishes the context of the work being reported;

iii. Discusses the related primary research literature (with citations) and

iv. Summarizes current understanding of the problem under investigation;

v. States the purpose of the work in the form of the theory, question, or problem investigated; and,

vi. Briefly explain rationale behind each step;

vii. The approach and, whenever possible, the outcomes the study can reveal.

viii. Information that allows reader to fully understand the paper topic; the topic’s relevance; and the paper’s thesis before proceeding into more in-depth examination or exploration.

There are few more terms describing discourse relations of the scientific research papers but we have not added them for our search as these terms will provide repetitive already extracted or
identified sentences. These terms may be we illustrate; more precise; in addition; hence; it concerns; provided that; however; as for; let us consider; regarding; first of all; it is necessary to emphasize; we would assume; it may be admitted; we call it/ them; on the one hand; as illustrated below; for example; next; finally; in our opinion; it seems reasonable; summing up etc.

4.4 System Architecture

This innovative system is named as Optimized Summarization System (‘OSS’). The OSS algorithm works on single as well as multiple research papers ‘Abstract’ and ‘Introduction’ sections only. These retrieved papers are mostly available in PDF, HTML or post-script formats. These papers then converted into plain text formats and only first two sections are segmented for summarization. This approach first generates summary of each individual document and then same similarity comparison takes place between summaries of individual document to produce multi-document summary.

4.4.1 Generic Multi-document Summarization

The generic architecture of summarization systems consists of basically three modules such as text preprocessing, summarization algorithm, and post-processing as shown in figure 4.3. The preprocessing module performs all the tasks of cleaning or making input text ready to be supplied for summary generation algorithm. It basically handles the activities such as sentence filtering, removing unnecessary contents from documents, indexing, tokenization, lemmatization, stemming etc. Once the input text is preprocessed it is supplied to summarization algorithms for summary generation. These summarization algorithms are the innovative ways of producing summaries depending on generic or user specific needs. Once the summary will be generated it is post processed and finally user expected summary is produced.

4.4.2 OSS Architecture

The innovative ‘Optimized Summarization System’ architecture is shown in figure 4.4. The input documents, research papers are loaded and preprocessed as a text by text preprocessing module which annotates the text with additional textual information [108]. These text preprocessing modules include a tokenizer and a stopword remover. RRN terms are then identified by term identifier
module. Research related scholar expected categories are generated using research category generation module. Category-wise related sentences are extracted from papers using sentence extraction module [108]. It is followed by sentence selection and sentence scoring modules in order to assign an important score to each sentence.

![Diagram](image)

**Figure 4.3:** Generic Multi-document Summarization Architecture

The sentence scoring modules use the information added by the text processing modules. The sentence extraction component takes the scores assigned to sentences and produces the summary of a text. In the cases where the already extracted sentences are taken into consideration before extracting more sentences, the information about extracted sentences is fed back onto the RRN identification module and used by the appropriate sentence scoring module to rescore sentences which were not yet extracted. Thus the redundancy is achieved through optimization.
Figure 4.4: Innovative OSS Architecture
4.5 OSS Algorithm

The Optimized Summarization System algorithm basically consists of six steps as follows:

Step1. Preprocessing the research papers i.e. (Preprocessing).

Step2. Identifying the RRN terms and categorize them i.e. (RRN Term Identification and categorization).

Step3. Measuring the similarity between sentences and clustering them i.e. (Similarity Measure and Sentence Clustering).

Step4. Selecting representative sentences and scoring them i.e. (Sentence Selection and Sentence Scoring).

Step5. Creating individual paper’s summary i.e. (Single Paper Summary).

Step6. Generating final multiple paper’s summary i.e. (Final Multi-Paper Summary).

Step1. Preprocessing:

The papers are preprocessed before inputted to the system. This step basically divided into two activities, text formatting and morphological processing / word processing.

A. Formatting - This indicates how the information is encoded in the source is one of the input characteristics of summarization. The increase in the number of documents available in formats such as PDF, HTML and Postscript, it is necessary to encode the source in a format which needs to be converted into a more convenient one in order to be processed.

B. Morphological processing - This factor is important because if no processing is applied before a word is scored, two different morphological forms of the same root are considered to be completely different concepts, and therefore two separate scores are calculated for each one. As a result, the score received by the concept denoted by these morphological forms will not be correct. This processing is done through identifying words, tokenization and removing stopwords.

The preprocessing involves the following steps:

The language for all the input sources is English, as is the output.
i. Segmenting plain text file into ‘Abstract’ and ‘Introduction’ section sentences only.
ii. Identify and remove non textual information such as formulas, tables, figures, eventual LATEX mark ups and citations.
iii. Splitting sections into sentences by sentence boundary detection (e.g. “.”, “?” “!”). Sentences in turn split into words.
iv. Tokenization consisting stemming, removal of stop/noisy words, punctuation from indexed data.

   a. *Stemming* - reduces a word to its stem using a set of predefined rules. The stemmer employed here is the Porter stemmer [29] which relies on a set of rules to remove affixes. For the two words mentioned above, the stemmer will identify the stems *comprehends* and *compress* respectively. Stemming is usually employed in information retrieval to match the query terms with the text in a document because it is very fast, and in most cases the results returned are acceptable.

   b. *Stopword Remover* - Not all the words are useful for scoring purposes. For example, prepositions such as a, an, the, in, by etc., appear very frequently in texts, but they provide little information about the importance of a sentence. The most common approach of dealing with these words is to build a list of *stopwords* called a *stoplist*, and to filter out these words during the scoring process. These words can be articles and prepositions, but also frequent verbs, adjectives, conjunctions, pronouns and auxiliary verbs and adverbs. Stopwords which can be readily used [71] and removed are available online.

In this way text is cleaned for automatic processing. These activities are considered to be the preliminary steps in summary generation to skim and scrutinize the documents.

**Step2. RRN Term Identification and Categorization:**

i. Identifying sentences related to the query using cosine similarity and term identification with a threshold below which no sentences will be selected.
ii. Using segmentation tools a sentence is split into a list of “words”.
iii. Identifying individual words or phrases \(W\), reflecting their significance in the text and having relative complete meaning named as RRN terms.
iv. Dividing all sentences from ‘Abstract’ and ‘Introduction’ sections into five categories as research goals, research methods, similarities or contrast, continuation or novel, and research outcome, which are words or phrases representing theses categories (Wtype) as term types.

Few terms/phrases which are commonly used in ‘Abstract’ and ‘Introduction’ sections representing all research concentrated ‘novelty’ presentations i.e. RRN terms as shown in figure 4.2 are identified manually by studying most of the scientific research papers. All the sentences starting with or containing these common terms are extracted. We are not concerning the length of the sentence while extraction. In [9, 55] one of the features of a machine learning method, identifies sentences with less than 5 words. The assumption here is that if a sentence is too short, it is less likely to include important information and is not worth including in a summary.

**Step3. Similarity Measure and Sentence Clustering:**

The procedure for finding similar sentences from preprocessed documents to cluster them into groups, we choose the maximal marginal relevance (MMR) [67] metric approach because of its simplicity and effectiveness. It is not a statistical learning approach and does not require any training data. It extracts the most related sentences and at the same time avoids redundancy in the summaries. The MMR metric is defined as below.

i. Depending on the desired length of the summary, select a few or a large number of sentences to compute sentence redundancy using the cosine similarity metric

ii. Used sentence similarity scoring as method of clustering sentences.

iii. Users can select the number of sentences or the amount of compression.

MMR metric is defined as,

\[
MRR(P, C, Q, R, S) = \arg \max_{P_{ij} \in R \setminus S} \left[ \lambda \cdot Sim_1(P_{ij}, Q, C_{ij}) - (1 - \lambda) \cdot \max_{P_{nm} \in S} (Sim_2(P_{ij}, P_{nm}, C, S)) \right]
\]

(1)

Where,

D = document collection

P = sentences from document collection (e.g. \( P_{ij} \) is the sentence j from document \( D_i \))

C = set of sentence clusters for the set of documents D
OPTIMIZED SUMMARIZATION OF RESEARCH PAPERS USING DATA MINING STRATEGIES

\( C_{ij} \) = subset of clusters of \( C \) that contains sentences \( P_{ij} \)
Q = query/topic specification
R = IR (D, P, Q, \( \Theta \)) = ranked list of sentences from documents retrieved by IR system, where \( \Theta \) = relevance threshold, below which it will not retrieve any sentences. (\( \Theta \) = can be degree of match/number of sentences)
S = subset of sentences in R already selected
R\( \setminus \)S = set difference i.e. set of sentences in R not yet selected

To compute standard relevance ranked list plus some additional scoring factors set \( \lambda \) =1.
To compute maximal diversity ranking among the documents in R set \( \lambda \) =0.
For intermediate values of \( \lambda \) between [0, 1] a linear combination of both criteria is optimized.

iv. Calculate \( \text{Sim}_1 \), similarity metric for relevance ranking as:

\[
\text{Sim}_1 = (\text{cosine similarity metric of sentence and query} + \text{coverage score for the sentence by whether the sentence is in one or more clusters and the size of the cluster} + \text{information content of the sentence by taking into account the RRN terms (figure 4.2)).}
\]

(2)

Cosine similarity measurement is very common way of calculating corpus based sentence similarity or string similarity. This metric is the identification of maximum relevance between sentences from document collection. The cosine sentence similarity is calculated as [73];

\[
\text{sim}(S_i,S_j) = \cos(T_i,T_j) \frac{\sum_{i=1}^{n} t_{1i}t_{2i}}{\sqrt{\sum_{i=1}^{n} t_{1i}^2 \cdot \sum_{i=1}^{n} t_{2i}^2}}
\]

(3)

Where \( S_i \) and \( S_j \) are two sentences and \( t_i \) is the term weight.

v. Calculate \( \text{Sim}_2 \), similarity metric for anti-redundancy as:

\[
\text{Sim}_2 = (\text{cosine similarity metric of sentence and previously selected sentence} + \text{penalize sentences that are part of clusters from which other sentences have already been chosen} + \text{penalize documents from which sentences have already been selected});
\]

(4)

A value of 1 means there is a perfect match between extracted sentence and the query. The similarity score is based on the centroid value of a sentence obtained from sentence clusters.
which only considers those words with a high TF.IDF (term frequency and inverse document frequency) score (step 4). TF*IDF has been widely used in information retrieval [72, 73, 74] and at present it is the scoring method employed by most of the term-based summarization methods [55, 56].

The clustering approach is effective in text summarization domains, where the features are asymmetric binary and hence cluster centroids are a meaningful description of the clusters [75]. Number of clusters i.e. the number of sentences in the desired length summary is an acceptable value for the number of clusters. The most probable number of sentences in a fixed-length-summary will be the length of summary fixed by user divided by the average length of sentences in ‘Abstract’ and ‘Introduction’ document collection. Identifying the number of optimal clusters, OSS uses K-Means clustering algorithm [108]. Each of the output sentence clusters is supposed to denote one research category in the document collection.

**Step4. Sentence Selection and Sentence Scoring:**

The main component of term-based summarization is the term scoring method because it is the basis on which each term in a text is scored, which in turn used to compute the score of a sentence. Selecting representative sentences from the clusters is a key problem. In this research, the sentence selection decision is determined not only by the content of the sentence, but also by the rest of the sentences extracted. For each sentence cluster, one sentence is selected to represent the category denoted by the cluster.

i. Each sentence denotes the Document ID number, Sentence Number to be needed by final summary.

ii. The RRN terms extracted from the papers are supposed to denote the research categories in the papers.

iii. Weight the sentences based on the RRN terms included in the sentences.

iv. Use three features to select the representative sentence: centroid sentence, TF*IDF and Term Frequency.

v. Use the RRN terms extracted from the clusters, to determine the focus or research category of each cluster.
i) **Centroid Sentence** - Centroid sentence is selected by two steps. First, the centroid vector of the cluster is calculated. Second, the sentence, which has the smallest distance with the centroid vector, is selected.

ii) **TF*IDF (Term Frequency*Inverse Document Frequency)** - Among a cluster, the sentences with the highest scores are selected as the representative sentences. For this TF*IDF score is used. The term frequency, \( tf(t, d) \) is the occurrence count of a RRN term ‘t’ in a research papers document ‘d’ with paper collection ‘D’.

\[
\text{tf}(t_{i,j}, d) = \text{raw frequency of the RRN terms at position } i,j \text{ in document } d. \quad (5)
\]

Using only term frequency does not consider the weight of a term in a document. Some terms are simply more important than others due to their rarity of use, do a better job of identifying the related sentences. Therefore the inverse document frequency which is a measure of the general importance of the terms from research paper document collection is also calculated. The \( IDF(t) \), inverse document frequency is given by,

\[
IDF(t) = \log(|D|/|\{d: t \in d\}|) \quad (6)
\]

where ‘D’ is the number of documents in the collection, and ‘d’ is the research paper in the collection which contain the RRN term ‘t’. Since \( IDF \) on its own is a relatively weak indicator of the term’s importance and for this reason very often it is used in conjunction with the term frequency (TF) as \( TF - IDF(t, d) \).

\[
TF - IDF(t, d) = tf(t, d) \times idf(t) \quad (7a)
\]

is nothing but,

\[
TF - IDF(t, d) = tf(t, d) \times \log(|D|/|\{d: t \in d\}|) \quad (7b)
\]

For sentence scoring in addition to sentence selection, term weighting score are also important. In research papers the terms repeated most are linked to the main topics [3]. Thus the Term Weight (\( W_{ij} \)) is the number of total occurrence of a RRN term in collection with number of
sentences; it is the normalized term frequency in the sentence collection. Each sentence is represented as the weights of terms. The influence of different term weighting methods has been extensively investigated in information retrieval, where researchers tried to find out the best term indexing method [38, 76, 77, 78].

Using these observations, it is possible to assign a score to each RRN term which is equal to its frequency in order to indicate the research category represented by it.

\[
Score (t) = tf(t, d) = W_{ij}
\]

This scoring technique was the first employed to produce automatic summaries [3, 4] and has the advantage that it can be easily calculated. Once each sentence is scored those sentences are ranked based on the descending order of weights.

**Step5. Single Paper Summary:**

Once the important information is located, sentences are extracted in their original order and form to form the summary. This information is presented into five research categories. The raw score of each sentence is calculated in above step and is compared between multiple sentences from each research category. Score comparison can be arranged in ascending or descending order. User will then select the threshold i.e. number of sentences per research category to form individual document summary. Thus the sentences with maximum and or minimum weight will be included or excluded from single paper summary document respectively.

**Step6. Final Multi-Paper Summary Generation:**

To view the summary of multiple papers, single document summaries of individual research papers from different clusters are combined together. Thus the scholar can compare between multiple single paper summaries presented as multi paper summary. In order to generate desired percentage of summary, a threshold is set as:
Threshold = ((Total sentences of first document) + (Total sentences of second document) + 
(Total sentences of third document) +...+ (Total sentences of n\textsuperscript{th} document)) × (Desired 
percentage of summary). \hspace{1cm} (9)

Thus for final multi-paper summarization, the elected sentences are stored till the desired 
percentage for summarization is met for each single paper summary under various categories. By 
comparing multiple papers’ query topic, now the scholar can decide on which document should 
be selected for further in depth reading.