Chapter 7
Results and Discussion

Overview:

The ‘Optimized Summarization System’, OSS identifies innovation through ‘Research Related Novel Terms’ (RRN) based analysis which presents the original contribution of this thesis. The scientific research papers are summarized under various research categories stating innovative scientific ideas presented in papers (research aim/goal), uniqueness and difference from previous published papers (research similarity or dissimilarity), research approach & methodology used for implementation (research methods/approach), research continuation of earlier/existing work (research continuation/novel) and research results & discussions published (research outcome).

7.1 Overall Results

In the information retrieval literature, efficiency is used to refer how “fast” a system retrieves documents and effectiveness is used to refer how “relevant” the documents are in relation to the query. Use of F-measure [88] is a convenient way for reporting precision ($P$) and recall ($R$) in one value. Precision is the ratio of relevant retrieved documents to retrieved documents while recall is the ratio of related retrieved documents to relevant documents.

To determine the relevance of each paper, a four-point scale was used to calculate precision as follows:
A research paper document representing all five categories of RRN terms is given a score of three.

A research paper document representing only four categories of RRN terms is given a score of two.

A research paper document representing only three categories of RRN terms is given a score of one.

A research paper document representing other than the above (i.e. only two or one category) is given a score of zero.

A research paper document occurring more than once under different URL (formats) is assigned a score of zero.

A non response of the paper document for all RRN categories is assigned a score of zero.

Then the \textit{precision} is calculated as:

\[
\text{Precision} = \frac{\text{Sum of the scores of RRN category for research documents retrieved}}{\text{Total number of Research documents evaluated}}
\]  
(12)

And the relative \textit{recall} value is thus defined as:

\[
\text{Recall} = \frac{\text{Total number of RRN categories for research documents retrieved}}{\text{Sum of research documents retrieved by all five categories}}
\]  
(13)

\textit{F-measure} \cite{41} is a convenient way for reporting \textit{precision} and \textit{recall} in one value. For evaluating performance average across categories and overall performance scores, summaries are first determined by computing performance measures per category then averaging these to compute the global means. The \textit{F-measure} is calculated as,

\[
F - \text{ measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}
\]  
(14)
OPTIMIZED SUMMARIZATION OF RESEARCH PAPERS USING DATA MINING STRATEGIES

### Research Goals

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS</td>
<td>52</td>
<td>44</td>
<td>65</td>
<td>26</td>
<td>34</td>
<td>20</td>
<td>61</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>HUMAN</td>
<td>63</td>
<td>72</td>
<td>56</td>
<td>52</td>
<td>50</td>
<td>55</td>
<td>79</td>
<td>79</td>
<td>79</td>
</tr>
</tbody>
</table>

### Research Methods

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS</td>
<td>26</td>
<td>34</td>
<td>20</td>
<td>57</td>
<td>66</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>66</td>
</tr>
<tr>
<td>HUMAN</td>
<td>79</td>
<td>79</td>
<td>79</td>
<td>79</td>
<td>79</td>
<td>93</td>
<td>94</td>
<td>92</td>
<td>86</td>
</tr>
</tbody>
</table>

### Research Similarity/Dis similarity

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>66</td>
<td>57</td>
<td>66</td>
<td>86</td>
<td>84</td>
<td>88</td>
</tr>
<tr>
<td>HUMAN</td>
<td>93</td>
<td>94</td>
<td>92</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>71</td>
<td>68</td>
<td>75</td>
</tr>
</tbody>
</table>

### Research Continuation/Novel

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS</td>
<td>66</td>
<td>66</td>
<td>88</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>66</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>HUMAN</td>
<td>86</td>
<td>84</td>
<td>88</td>
<td>88</td>
<td>84</td>
<td>88</td>
<td>71</td>
<td>68</td>
<td>75</td>
</tr>
</tbody>
</table>

### Research Outcome

<table>
<thead>
<tr>
<th></th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
<th>F</th>
<th>P</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>OSS</td>
<td>52</td>
<td>44</td>
<td>65</td>
<td>26</td>
<td>34</td>
<td>20</td>
<td>61</td>
<td>57</td>
<td>66</td>
</tr>
<tr>
<td>HUMAN</td>
<td>52</td>
<td>50</td>
<td>55</td>
<td>79</td>
<td>79</td>
<td>79</td>
<td>93</td>
<td>94</td>
<td>92</td>
</tr>
</tbody>
</table>

**Table X**: Performance per category: *F*-measure (*F*), precision (*P*) and recall (*R*).

Table X shows the performance per research categories for human verses OSS with *F*-measure (*F*), precision (*P*) and recall (*R*).

Table XI shows the percentage *F*-measure score for human produced and automated OSS with all research categories.

Figure 7.1 and 7.2 shows the comparison graph of performance per category: *F*-measure for human produced summaries and automated OSS.

### Research Categories

<table>
<thead>
<tr>
<th>Research Categories</th>
<th><em>F</em>-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(OSS Summary)</td>
</tr>
<tr>
<td>Research Goals</td>
<td>62%</td>
</tr>
<tr>
<td>Research Methods</td>
<td>51%</td>
</tr>
<tr>
<td>Research Similarity/Dissimilarity</td>
<td>79%</td>
</tr>
<tr>
<td>Research Continuation/Novel</td>
<td>92%</td>
</tr>
<tr>
<td>Research Outcome</td>
<td>71%</td>
</tr>
</tbody>
</table>

**Table XI**: Percentage *F*-measure Score for Human generated and Automated OSS Summaries.
Figure 7.1: Performance per category: $F$-measure for OSS.

Figure 7.2: Performance per category: $F$-measure for Human Summary.

From table XI and figure 7.1 and 7.2, it is clear that the OSS obtains substantial improvement over the human in terms of precision and recall of the important categories such as Research Goals, Research Methods and Research Outcome.
7.2 Experimental Results

OSS experimentations were conducted using open source Derby Database and Java language.

![Flow Diagram for OSS execution](image)

**Figure 7.3**: Flow Diagram for OSS execution
Derby’s database format is platform independent allowing databases to be copied to any machine type. The results show that OSS identifies innovation through ‘Research Related Novel’ term analysis and presents the main research contributions of the published research papers. The flow of experimentation steps are shown in figure 7.3 followed by some of the examples of OSS output.

Figure 7.4 shows free open source Derby Database tool installation procedure which enables Derby environment driver.

```
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.
C:\Users\Mrs. Sunita Patil>cd
C:\>cd C:\\db-derby-10.8.2.2-bin\\bin
C:\db-derby-10.8.2.2-bin>set EmbeddedCP.bat
C:\db-derby-10.8.2.2-bin>set DERBY_INSTALL=C:\db-derby-10.8.2.2-bin
C:\db-derby-10.8.2.2-bin>set DERBY_HOME=C:\DB-DER~1.2-B
C:\db-derby-10.8.2.2-bin>set CLASSPATH=C:\DB-DER~1.2-B\lib\derby.jar;C:\DB-DER~1.2-B\lib\derbytools.jar;C:\db-derby-10.8.2.2-bin\lib\derby.jar;C:\db-derby-10.8.2.2-bin\bin
C:\db-derby-10.8.2.2-bin>
```

**Figure 7.4: Enabling Derby Environment Driver**

Figure 7.5 shows the procedure needed to connect Derby database tool with OSS executable file through java database ‘jdbc’ connectivity where ‘ij’ is Derby’s interactive ‘jdbc’ scripting tool.

```
c:\db-derby-10.8.2.2-bin>ij.bat
ij version 10.8
ij> connect ’jdbc:derby:wba’;
ij>
```

**Figure 7.5: Connectivity with Database**

Figure 7.6 to 7.10 shows examples of Research Relevant Novel term creation for various research categories. We identified various RRN terms under categories such as Research Goals, Research Methods, Research Continuation or Novel, Research Similarity/Dissimilarity and Research Outcomes.
the aim of this
we observe
this paper
study the problem of
research goal
address the issue of
consists of
the purpose of
.
.
30 rows selected

**Figure 7.6: RRN terms for Goal Category**

mechanics suggests
new method
be use to estimate
method
scheme
new way
architecture
.
.
18 rows selected

**Figure 7.7: RRN terms for Method Category**

system extend to
improve the performance of
in our first experiment
we use the framework
more closely related to
our recent work
other researchers claim that
earlier
similar work
previous work
novel

33 rows selected

**Figure 7.8**: RRN terms for Research Continuation/Novel

```sql
ij> select * from contrast_like;
CL
------------------------
contrast to
unlike
existing
like
prior work by
as opposed to
in comparison to
previous study

33 rows selected
```

**Figure 7.9**: RRN terms for Similarity/Dissimilarity

```sql
ij> select * from outcome;
OUTCOME
------------------------
prove that
at the end of
the result show that
believe that
output
result indicate
conclude
performance
```
experiment demonstrate that
result carried out
the evaluation
outcome
prove that

30 rows selected

Figure 7.10: RRN terms for Outcome Category

Figure 7.11 shows the execution of OSS tool. First it will ask number of input documents for summarization. The user will then insert number of research documents to be summarization.

C:\db-derby-10.8.2.2-bin\bin>javac Wba.java
C:\db-derby-10.8.2.2-bin\bin>java wba
   OPTIMIZED SUMMARIZATION SYSTEM (OSS)

NO. OF PAPERs TO SUMMARIZE:
1

Fetching Database......

Figure 7.11: Number of Documents to summarize

The fetching of database starts which compares RRN terms from various categories with terms found in input documents. This is shown in figure 7.12.

[GOAL RRN TERMS]

[the aim of this, we observe, this paper, study the problem of, research goal, address the issue of, the characteristic of, be find that, consists of, the meaning of, have prove, is depicted as, understand the problem of, in the domain of, more specifically to, the purpose of, paper explore, paper provide, find solution, be propose, is meant for, is designed for, give an overview of, study show at, the purpose]
Figure 7.12: Fetching Database

After fetching the database sentence boundary identification module gets executed, this identifies the boundary or end of each sentence. This also gives total number of lines in Abstract and Introduction sections under consideration as shown by figure 7.13 below.

Number of Lines in PAPER = 19

How to effectively manage spectrum data to find information oriented subjects?
How to mine the spectra data in observation intervals from data in existence?
In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.

A new method on mining canopy a spectrum in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.

At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, China.

Hyperspectral remote sensing comes into the application stage after twenty years development, which gives us abundant spectral information of the ground objects.

So, institutes for basic remote sensing research in many countries invest a lot of manpower and material resources into collecting all kinds of spectra.

Database on spectra of ground objects come into being in bulk, such as spectra database of USGS, database on standard substance component of JPL and spectra database of John Hopking University etc.

**Figure 7.13: Sentence Boundary Identification**

The sentence extraction data mining module compares each sentence against research categories and collects them under various clusters as shown in figures 7.14 and 7.15 for Goal and Method clusters respectively.

**SENTENCE EXTRACTION MODULE**

**GOAL OF PAPER:**

1. In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.

2. In this paper, data warehouse is introduced into organizing spectra data.

3. A new method about canopy spectra data mining based on crop models is given in this paper whose feasibility is tested by a case.

**Figure 7.14: Category-Goal Sentence Extraction**

**SENTENCE EXTRACTION MODULE**

**METHOD OF PAPER:**

1. A new method on mining canopy spectra in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.
2. At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, China.

3. A tree data structure containing a mapping relation between fact tables and dimensional tables of multidimensional tables is presented, by which user can query spectra and its matching parameters together.

4. A tree data structure is created in order to organized data oriented vegetation spectra.

5. But there are still some problems about uncertainty of simulation, deeply matching of data and integrality of tree data structure that are to be done by our next work.

6. This work prove that it is better than other systems.

**Figure 7.15: Category-Method Sentence Extraction**

Figure 7.16 shows the result of tokenization module which separates tokens from extracted sentences for similarity check and relevance judgment.

**Figure 7.16: Tokenization**

The weight of every extracted sentence is calculated depending on number of research terms available in it assigning a score to it and similarity between sentences are measured by using cosine similarity check. The same is shown in figure 7.17 and 7.18 for Goal and Method categories respectively.

**Figure 7.17: Goal-Sentence Scoring & Similarity Measure**

**STANDARD SENTENCE SCORING & SIMILARITY MEASURE**

**GOAL SENTENCE WEIGHT:**
1. In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.

   WEIGHT: 13

2. In this paper, data warehouse is introduced into organizing spectra data.

   WEIGHT: 3

3. A new method about canopy spectra data mining based on crop models is given in this paper whose feasibility is tested by a case.

   WEIGHT: 5

ENTER SUMMARY THRESHOLD (not more than 3):
SENTENCE SCORING & SIMILARITY MEASURE

METHOD SENTENCE WEIGHT:
1. A new method on mining canopy spectra in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.
   WEIGHT: 9
2. At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, Chin
   WEIGHT: 7
3. A tree data structure containing a mapping relation between fact tables and dimensional tables of multidimensional tables is presented, by which user can query spectra and its matching parameters together.
   WEIGHT: 10
4. A tree data structure is created in order to organized data oriented vegetation spectra.
   WEIGHT: 4
5. But there are still some problems about uncertainty of simulation, deeply matching of data and integrality of tree data structure that are to be done by our next work.
   WEIGHT: 2
6. This work prove that it is better than other systems.
   WEIGHT: 1

Figure 7.18: Method-Sentence Scoring & Similarity Measure

After measuring score and similarity between sentences the highest scored sentences will selected for final summary. The user can set a threshold value for in more or less summary sentences as shown in figure 7.19.

GOAL OF PAPER:
- In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.
- A new method about canopy spectra data mining based on crop models is given in this paper whose feasibility is tested by a case.

METHOD OF PAPER:
- A tree data structure containing a mapping relation between fact tables and dimensional tables of multidimensional tables is presented, by which user can query spectra and its matching parameters together.
- A new method on mining canopy a spectrum in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.
- At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data
mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, China.

RESEARCH_CONTINUATION OF PAPER:
- Canopy spectra data mining is studied based on crop models and data warehouse in order to gaining spectra in observation intervals and giving user a series of temporal spectra.

CONTRAST/LIKE OF PAPER:
- THERE ARE NO RELEVANT STATEMENTS IN CONTRAST/LIKE CLUSTER.

OUTCOME OF PAPER:
- But there are still some problems about uncertainty of simulation, deeply matching of data and integrality of tree data structure as results are identified.

SEE SUMMARIZED PAPER AT C:\db-derby-10.8.2.2-bin\bin\output.txt.

#### Figure 7.19: Category-wise Summary

The complete summary of a single or multiple documents can be stored as a text file for further readings. The single document summary is shown in figure 7.20 and multiple document summaries are presented in figure 7.21.

************** Paper-1-Summary **************

GOAL SUMMARY STATEMENTS:
- In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.
- A new method about canopy spectra data mining based on crop models is given in this paper whose feasibility is tested by a case.

METHOD SUMMARY STATEMENTS:
- A tree data structure containing a mapping relation between fact tables and dimensional tables of multidimensional tables is presented, by which user can query spectra and its matching parameters together.
- A new method on mining canopy a spectrum in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.
At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, China.

RESEARCH CONTINUATION SUMMARY STATEMENTS:

- Canopy spectra data mining is studied based on crop models and data warehouse in order to gaining spectra in observation intervals and giving user a series of temporal spectra.

CONTRAST/LIKE SUMMARY STATEMENTS:

THERE ARE NO RELEVANT STATEMENTS.

OUTCOME SUMMARY STATEMENTS:

- But there are still some problems about uncertainty of simulation, deeply matching of data and integrality of tree data structure as results are identified.

Figure 7.20: Output File-Single Document Summary

<table>
<thead>
<tr>
<th>paper-1-summary</th>
<th>paper-2-summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOAL SUMMARY STATEMENTS:</td>
<td>GOAL SUMMARY STATEMENTS:</td>
</tr>
<tr>
<td>- In this paper, a tree data structure, which includes dimension tables and fact tables, is presented to establish the top architecture and tie together the various data tables, for instance, spectrum data and its interrelated matching parameters.</td>
<td>- In this paper popular Data Mining segmentation techniques are presented along with incremental learning algorithms, as well as a new segmentation method with the use of genetic algorithm.</td>
</tr>
<tr>
<td>- A new method about canopy spectra data mining based on crop models is given in this paper whose feasibility is tested by a case.</td>
<td>- The term Data Mining is mostly used in the context of searching for hidden dependencies in data, which existence does not have to be evident to the analyst.</td>
</tr>
</tbody>
</table>

METHOD SUMMARY STATEMENTS:

- A tree data structure containing a mapping relation between fact tables and dimensional tables of multidimensional tables is presented, by which user can query spectra and its

- The use of segmentation methods requires defining a distance function between any two
matching parameters together.

- A new method on mining canopy a spectrum in observation intervals is discussed, which is based on certain crop models and the tree structure data warehouse.

- At last, a case on vegetation spectral simulation is given to test structure of data warehouse and verify data mining method using SE590 data obtaining during the field experiment conducted during the year before last in Shunyi, Beijing, China.

**RESEARCH_CONTINUATION SUMMARY STATEMENTS:**

- Canopy spectra data mining is studied based on crop models and data warehouse in order to gaining spectra in observation intervals and giving user a series of temporal spectra.

**CONTRAST/LIKE SUMMARY STATEMENTS:**

**THERE ARE NO RELEVANT STATEMENTS.**

**OUTCOME SUMMARY STATEMENTS:**

- But there are still some problems about uncertainty of simulation, deeply matching of data and integrality of tree data structure as results are identified.

- The query for such dependencies cannot be simply expressed in the query language, and it requires the use of advanced algorithms of data processing.

- These techniques can be divided into a few basic categories, with respect to the kind of data processing they perform.

**RESEARCH_CONTINUATION SUMMARY STATEMENTS:**

**THERE ARE NO RELEVANT STATEMENTS.**

**CONTRAST/LIKE SUMMARY STATEMENTS:**

- In a modern data warehouse with incremental data update, at every change there is only a small amount of new data present, however most of the Data Mining techniques requires training the model on the full training set.

- However extracting the aspect of interest from the large amount of data is usually not a trivial task.

**OUTCOME SUMMARY STATEMENTS:**

- The presented results show that k-means segmentation with genetic algorithm can dynamically adapt the value of k to the data.

- Standard k-means in case of scattered data was giving particularly bad results.

**Figure 7.21:** Output File-Multiple Document Summary

### 7.3 Discussion on the Results

The figure number 7.1 to 7.21 demonstrates the architecture based OSS implementation outputs. All the OSS modules are executed with sample input documents.
These results inform the following achievements gained by implementation of OSS:

- The TF*IDF ranking, term frequency ranking produces similar results, while the centroid sentence ranking produces the summaries with lower performance. The reason may be that the information of terms frequency, their cluster frequency, as well as their length, is useful to select better representative sentences.
- These summarization results also show that this system is an effective and efficient strategy for providing short, condensed, accurate, explicit, optimized and most related multiple research paper’s summary, minimizing research scholar’s efforts deciding whether to go ahead with the retrieved papers for further readings and or references helping in his/her own work.
- This hypothesis was tested with the results of several researchers as they report very different levels of reuse.
- In [9] it is found that 79% of the sentences in the abstracts could be perfectly matched with sentences in the full text, whereas [102] observed that only 31.7% of the sentences in the abstracts can be matched.
- Thus the percentage of matching is even lower for the experiment presented in [103] where only 15% of the clauses could be paired with clauses from the full text.
- Even though the results presented in these papers are of very different which are domain specific, the authors noticed that in many cases where no direct match is found, it is possible to identify several sentences from the document which were combined to produce a sentence in the abstract.