

DESIGN OF NOVELTY DETECTION TECHNIQUES FOR OPTIMIZED SEARCH ENGINE RESULTS

THESIS

submitted in fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

J.C. BOSE UNIVERSITY OF SCIENCE AND TECHNOLOGY, YMCA

by

SUSHIL KUMAR

Registration No: YMCAUST/Ph.D-08/2012

Under the Supervision of

Dr. KOMAL KUMAR BHATIA

PROFESSOR



Department of Computer Engineering

Faculty of Engineering & Technology

J.C. Bose University of Science and Technology, YMCA

Sector-6, Mathura Road, Faridabad, Haryana, INDIA

JULY 2021

DECLARATION

I hereby declare that this thesis entitled **“DESIGN OF NOVELTY DETECTION TECHNIQUES FOR OPTIMIZED SEARCH ENGINE RESULTS”** by **SUSHIL KUMAR**, being submitted in fulfillment of requirement for the award of Degree of Doctor of Philosophy in the Department of Computer Engineering under Faculty of Engineering and Technology of J.C. Bose University of Science and Technology YMCA, Faridabad, during the academic year 2020-2021, is a bonafide record of my original work carried out under the guidance and supervision of **DR. KOMAL KUMAR BHATIA, PROFESSOR, DEPARTMENT OF COMPUTER ENGINEERING, J.C. BOSE UNIVERSITY OF SCIENCE & TECHNOLOGY, YMCA, FARIDABAD** and has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any degree either in this University or in any other University.

(SUSHIL KUMAR)

Registration No. YMCAUST/Ph.D-08/2012

CERTIFICATE

This is to certify that this thesis entitled “**DESIGN OF NOVELTY DETECTION TECHNIQUES FOR OTIMIZED SEARCH ENGINE RESULTS**” by SUSHIL KUMAR submitted in fulfillment of the requirements for the award of Degree of Doctor of Philosophy in **DEPARTMENT OF COMPUTER ENGINEERING**, under Faculty of Engineering and Technology of J.C. Bose University of Science and Technology, YMCA, Faridabad, during the academic year 2020-21, is a bonafide record of work carried out under my guidance and supervision.

I further declare that to the best of my knowledge, the thesis does not contain any part of any work which has been submitted for the award of any degree either in this university or in any other university.

Dr. Komal Kumar Bhatia

Professor

Department of Computer Engineering,

Faculty of Informatics & Computing,

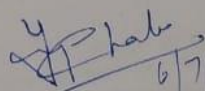
J.C. Bose University of Science & Technology, YMCA, Faridabad

Dated:

The Ph.D viva-voce examination of Research Scholar Sushil Kumar (YMCAUST/Ph.D-08-2012) has been successfully held on 06/07/2021.

(Signature of Supervisor)

(Signature of Chairperson)


6/7/2021
(Signature of External Examiner)

ACKNOWLEDGEMENT

I express my gratitude to almighty God for giving me strength and courage to complete this thesis.

I would like to express my sincere and deep gratitude to my Guru Dr. Komal Kumar Bhatia Professor and Dean of Faculty of Informatics and Computing, J.C Bose University of Science and Technology, YMCA, Faridabad for giving me the opportunity to work in this area. It would never be possible to take this thesis to this level without his innovative ideas, invaluable guidance, continuous support and encouragement. His knowledge of different perspectives of research provided me with the opportunity to broaden my knowledge and to make significant progress. At times when I got stumbled upon big obstacles, my mentor encouraged me to look further and keep sailing through tough times.

I gratefully acknowledge Dr. Komal Kumar Bhatia, Chairman Department of Computer Engineering for his constant encouragement and moral support during the course of this research work. I want to express my special thanks to Dr. Payal Gulati, Dr. Harish Kumar, Dr. Shailender Gupta, Ms. Sangeeta Dhell, Dr.Vedpal and Dr.Umesh Kumar. I gratefully acknowledge my university colleagues for their encouragement, support and invaluable suggestions in completing this research. I am also thankful to my students who helped me directly and indirectly in completing my research work.

Words fail to express my heartfelt thanks to my better half Ms. Shalini Panwar for the utmost patience, love and faithful support and unparallel availability at all times during the course of my work. A special thanks to my parents Sh. Bal Krishan Panwar and Smt. Sita Devi for all their love and encouragement throughout my education. I am also thankful to my Father-in-law Sh. Satypal Mogha, my mother-in law Smt. Santosh Mogha and Sh. Omkar Mogha for their blessings and support.

My special thanks to my son, Master Akshit Sushil Panwar for understanding me and giving me time for doing my research work.

Thank you all!

(Sushil Kumar)

ABSTRACT

World Wide Web is a large hyperlinked information reservoir which includes text, audio, video and metadata information etc. It has evolved as an important information resource for data available for all. Search Engines provide an interface to access information from this large information resource. Users can provide search query through search engines and the results, therefore are displayed on the screen in a ranked manner. Search engines maintain the web documents in indices and provide search facilities by continuously downloading Web pages for processing. This process of downloading of web pages is known as web crawling. In order to cope up with the increasing size of the web, search engines run many processes in parallel which are referred as parallel crawlers. Since multiple processes execute in parallel, they put tremendous pressure on network bandwidth. Overlap problem of web pages also occurs due to the reason that one crawling instance may not be aware of another already downloading the same page. Thus, the search engines provide a list of redundant documents which demand the optimization of search engines results. In this thesis novelty detection techniques to detect the novel information in text documents are proposed. These techniques provide relevant and novel information to the users by filtering out the redundant information, which results in less effort of the user in searching the new information.

A novel document-to-sentence level technique is proposed that splits the document into sentences on which the novelty is evaluated rather than the whole document. A threshold value is computed which is compared with the novelty score of the document, if this score is above the threshold value then the document is regarded as novel otherwise as not novel.

Extractive text summarization based novelty detection technique is also proposed in which the documents is summarized first on which the novelty is calculated. This calculated novelty is compared with the general novelty detection method. With the help of the implemented results it has been observed the proposed method provided better results in terms of redundancy removal.

Furthermore, Clustering based novelty detection method has been proposed for providing the novel results. A comparison is made between proposed method results and Bing search engine results. The proposed method provides better results in terms of elimination of duplicate documents and generating relevant and novel results to the user.

The work has also put forward the semantic similarity and text summarization base novelty detection technique. In addition the proposed system not only considers the synonyms for a given query keyword but also account all inferred word that are directly and indirectly related to the keyword for measuring content similarity. A comparison is made between generic crawler novelty and proposed method crawler novelty based on augmented hypertext documents. With the help of the implementation results it has been found that proposed work provides better results in terms of elimination of redundant documents with generating relevant results and novel documents identification to the user.

TABLE OF CONTENTS

Declaration	i
Certificate	ii
Acknowledgment	iii
Abstract	iv
Table of contents	vi
List of Tables	x
List of Figures	xii
List of Abbreviations	xv

CHAPTER I: INTRODUCTION 1-8

1.1	GENERAL	1
1.2	SEARCH ENGINES	1
1.3	MOTIVATION	2
1.4	NOVELTY DETECTION TO IMPROVE THE SEARCH EFFICIENCY	3
1.5	OBJECTIVES OF THE WORK	5
1.6	ORGANIZATION OF THE THESIS	7

CHAPTER II: LITERATURE SURVEY 9-22

2.1	INTERNET	9
2.2	WORLD WIDE WEB	10
2.3	INFORMATION RETRIEVAL	12
	2.3.1 Factors Affecting Performance in IR	12
2.4.	WEB DIRECTORIES	13
2.5	META SEARCH ENGINE	14
2.6	SEARCH ENGINE	14
2.7	NOVELTY DETECTION	14
	2.7.1 Event Level Novelty Detection	17
	2.7.2 Document Level Novelty Detection	17
	2.7.3 Sentence Level Novelty Detection	18
2.8	SUMMARY	22

CHAPTER III: DOCUMENT-TO-SENTENCE LEVEL NOVELTY DETECTION TECHNIQUE

25-48

3.1	INTRODUCTION	25
3.2	PROPOSED DOCUMENT TO SENTENCE LEVEL NOVELTY DETECTION	25
3.3	NOVELTY DETECTION ALGORITHM	25
3.3.1	Architecture of Proposed Novelty Detection Method	26
3.3.1.1	Segmentation Module	26
3.3.1.2	Text Categorization	27
3.3.1.3	Novelty Detector Module	27
3.3.1.4	Algorithm for Document Level Novelty Detection	30
3.4	SIMULATION OF THE PROPOSED APPROACH	31
3.4.1	Simulation	31
3.4.2	Example 1	31
3.4.3	Example 2	37
3.5	IMPLEMENTAION AND PERFORMANCE EVALUATION	41
3.5.1	Performance Evaluation	45
3.6	SUMMARY	48

CHAPTER IV: EXTRACTIVE TEXT SUMMARIZATION BASED NOVELTY DETECTION TECHNIQUE

49-66

4.1	INTRODUCTION	49
4.2	TEXT SUMMARIZATION	49
4.2.1	Types of Text Summarization	50
4.2.2	Extractive Text Summarization	50
4.3	TERM FREQUENCY	51
4.3.1	Keyword Frequency	51
4.4	INVERSE DOCUMENT FREQUENCY	51
4.4.1	Algorithm for TF-IDF	52
4.5	PROPOSED TEXT SUMMARIZATION BASED NOVELTY DETECTION	52
4.5.1	Algorithm for Proposed Technique	53
4.6	IMPLEMENTAION AND RESULTS	54
4.6.1	Experimental Result 1	54
4.6.1.1	Cosine Similarity Calculation	55

4.6.1.2	Apply the TF-IDF Technique on the Original Document	58
4.6.1.3	Apply the Cosine Similarity on the Summarized Data	59
4.6.1.4	TF-IDF plus Cosine Similarity Approach	61
4.6.2	Example 2	63
4.7	SUMMARY	66

CHAPTER V: CLUSTERING BASED NOVELTY DETECTION TECHNIQUE IN TEXT DOCUMENTS 67-80

5.1	INTRODUCTION	67
5.2	THE PROPOSED METHODOLOGY FOR CLUSTERING BASED NOVELTY DETECTION	67
5.2.1	Collection of Textual Datasets	68
5.2.2	Convert documents into Vectors	68
5.2.2.1	Tokenization	69
5.2.2.2	Stop-words Removal	69
5.3	NOVELTY DETECTION MODULE	69
5.4	IMPLEMENTATION OF CLUSERING BASED NOVELTY DETECTION	72
5.5	RESULT ANALYSIS	75
5.5.1	Performance Evaluation	78
5.6	SUMMARY	80

CHAPTER VI: SEMANTIC SIMILARITY AND TEXT SUMMARIZATION BASED NOVELTY DETECTION TECHNIQUE 81-105

6.1	INTRODUCTION	81
6.2	GENERIC CRAWLER METHODOLOGY	82
6.3	PROPOSED CRAWLER METHODOLOGY FOR NOVELTY DETECTION	84
6.3.1	Algorithm for Proposed Crawler Novelty Detection	86
6.3.2	Detailed Steps for Proposed Crawler Novelty	86
6.3.2.1	Sentence Parsing	87
6.3.2.2	Tokenization	87
6.3.2.3	Stop Words Removal	87

6.3.2.4	Noun Filtering Using word net 3.0	87
6.3.2.5	Word Overlap Value Calculation	88
6.3.2.6	Minimization of Data Using Ontology	88
6.4	SIMILARITY CALCULATION OF SUMMARIZED DATA	88
6.4.1	N-Gram Formation	88
6.4.2	Hash Conversion	89
6.4.3	Frame Parsing	90
6.4.4	Process of Fingerprints Selection	90
6.5	CALCULATION PROCESS OF DOCUMENT SIMILARITY	91
6.6	SIMULATION SET PARAMETERS	92
6.6.1	Set up Parameters	92
6.6.2	Performance Parameters	93
6.7	IMPLEMENATION	94
6.8	RESULTS AND DISCUSSION	96
6.9	SUMMARY OF RESULTS	105
CHAPTER VII: CONCLUSION AND FUTURE WORK		107-119
7.1	CONCLUSION	107
7.2	ENHANCED PERFORMANCE	108
7.3	FUTURE WORK	108
REFERENCES		111-119
BRIEF PROFILE OF RESEARCH SCHOLAR		121
LIST OF PUBLICATION		123-124

LIST OF TABLES

Table	Title	Page No.
Table 2.1	Methods for Novelty Detection	16
Table 3.1	Term and Frequencies of Sentence S1 for Calculating Cosine Similarity	28
Table 3.2	Term and Frequencies of Sentence S2 for Calculating Cosine Similarity	28
Table 3.3	Term and Frequencies of Sentence S1 and S2 together	29
Table 3.4	Cosine Similarity Values of N1 Document	35
Table 3.5	Cosine Similarity Values of N2 Document	35
Table 3.6	Cosine Similarity Values of N3 Document	35
Table 3.7	Select Maximum values from each table	36
Table 3.8	Novelty scores for N1, N2, N2	36
Table 3.9	Select Minimum Novelty Scores	36
Table 3.10	Cosine Similarity Values of M1, M2, M3	40
Table 3.11	Values of Novelty Scores	40
Table 3.12	Minimum Novelty Scores	40
Table 3.13	Comparison of Threshold Values	41
Table 3.14	Precision, Recall and F-Score Comparison	46
Table 3.15	Comparison with General Novelty Detection work	48
Table 4.1	TF values of Original Document	55
Table 4.2	Cosine values	56
Table 4.3	TFIDF values	58
Table 4.4	TF value after Text Summarization	59
Table 4.5	Cosine values	59
Table 4.6	TF values after TFIDF and Cosine similarity approach	61
Table 4.7	Threshold comparison	62
Table 4.8	TF and TFIDF values Example 2	64
Table 4.9	TF and TF-IDF Values after Summarization	65
Table 5.1	Distance matrix with Term frequency values	71
Table 5.2	Movement of clusters	72
Table 5.3	Comparison of Bing search Engine with Proposed Approach	76
Table 5.4	Precision, Recall and F-Score comparison	78

Table 6.1	Set up Parameters	92
Table 6.2	Comparison of General crawler and Proposed Crawler Novelty Data set-1	97
Table 6.3	Comparison of General crawler and Proposed Crawler Novelty Data set -2	100
Table 6.4	Comparison of General crawler and Proposed Crawler Novelty Data set -3	102

List of Figures

Figure	Title	Page No.
Figure 1.1	Components of Search Engine	2
Figure 2.1	Architecture of Internet	10
Figure 2.2	Basic Architecture of WWW	11
Figure 2.3	Complete Work Flow	23
Figure 3.1	Architecture of Proposed System	26
Figure 3.2	Novelty Detection Process	27
Figure 3.3	Algorithm for Proposed work	30
Figure 3.4	Example 1: document N1	31
Figure 3.5	Example 1: document N1	31
Figure 3.6	Example 1: document N3	32
Figure 3.7	Example 1: New document	32
Figure 3.8	Sentences of document N1	33
Figure 3.9	Sentences of document N2	33
Figure 3.10	Sentences of document N3	34
Figure 3.11	Sentences of New document	34
Figure 3.12	Example-2: Document M1	37
Figure 3.13	Example-2: Document M2	37
Figure 3.14	Example-2: Document M3	38
Figure 3.15	Example-2: New document	38
Figure 3.16	Sentences of Document M1	38
Figure 3.17	Sentences of Document M2	39
Figure 3.18	Sentences of Document M3	39
Figure 3.19	Sentences of New document	39
Figure 3.20	Query Interface	41
Figure 3.21	Input document for Processing	42
Figure 3.22	Document-2 (history document)	42
Figure 3.23	Generated Assumed Novelty Document	43
Figure 3.24	Novelty Score Computation	43
Figure 3.25	Maximum Cosine Similarity Value	44

Figure 3.26	Minimum Novelty Score Value	44
Figure 3.27	Novelty Detection Output	45
Figure 3.28	Precision Plot	46
Figure 3.29	Recall Plot	47
Figure 3.30	F-Score Plot	47
Figure 4.1	Algorithms for TF-IDF	52
Figure 4.2	Architecture for Text Summarization Based Novelty	53
Figure 4.3	Algorithms for Proposed Technique	54
Figure 4.4	Example 1: Original document	55
Figure 4.5	Extracted Novel Sentences	57
Figure 4.6	Document to apply TF-IDF method on Example-1	58
Figure 4.7	Summarized Documents after TF-IDF	58
Figure 4.8	Summarized Documents after Cosine Similarity	60
Figure 4.9	Extracted Novel Sentences after Cosine plus TF-IDF	63
Figure 4.10	Text Document Examples-2	63
Figure 4.11	Text Document2 after Preprocessing	64
Figure 4.12	Summarized Document-2	64
Figure 4.13	Document-2 for Proposed Technique	65
Figure 4.14	Extracted Novel Sentence	66
Figure 5.1	Architecture for Clustering Based Novelty Detection	68
Figure 5.2	Document chosen for Basic Analysis	69
Figure 5.3	Algorithm for K-means Clustering	70
Figure 5.4	Interface for Bing Search Engine	73
Figure 5.5	Module to train K-means model	74
Figure 5.6	Arrays for Clusters Head	75
Figure 5.7	Novel documents out of 30 documents	76
Figure 5.8	Novel documents out of 10 documents	77
Figure 5.9	Precision Plot	79
Figure 5.10	Recall Plot	79
Figure 5.11	F-Score Plot	79
Figure 6.1	Generic Web Crawler Architecture	82
Figure 6.2	Generic Crawler Novelty Interface	83
Figure 6.3	SQL Indexed Database	84

Figure 6.4	Semantic Similarity and Text Summarization based Novelty Detection Architecture	85
Figure 6.5	Ontology based text summarization	86
Figure 6.6	Frame parsing	90
Figure 6.7	Search Engine Interface	94
Figure 6.8	List of WebPages for the Query ‘code’ on Generic Crawler Search Interface	95
Figure 6.9	List of WebPages for the Query ‘code’ on Proposed Crawler Search Interface	96
Figure 6.10	Comparison of Generic Crawler and Proposed Crawler Novelty Results for Data Set-1	98
Figure 6.11	Comparison of Memory Overhead	99
Figure 6.12	Comparison of Generic Crawler and Proposed Crawler Novelty Results for Data Set-2	99
Figure 6.13	Comparison of Memory Overhead	101
Figure 6.14	Comparison of Generic Crawler and Proposed Crawler Novelty Results for Data Set-3	103
Figure 6.15	Comparison of Memory Overhead	104

LIST OF ABBREVIATIONS

WWW	World Wide Web
ARPANET	Advanced Projects Agency Network
DNS	Domain Name System
IP	Internet Protocol
NSF	National Science Foundation
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ICANN	Internet Corporation for Assigned Name and Number
W3C	World Wide Web Consortium
URL	Uniform Resource Allocator
IR	Information Retrieval
APWSJ	Associated Press World Street Journal
TDT	Topic Detection and Tracking
NED	New Event Detection
NID	New Information Degree
TF	Term Frequency
IDF	Inverse Document Frequency
SVM	Support Vector Machine
MMR	Maximal Marginal Relevance
LCA	Local Context Analysis
DND	Document Level Novelty Detection
SQL	Structure Query Language
CSV	Comma Separated Values
ASCII	American Standard Codes for Information Interchange
MD5	Message Digest 5
RR	Redundancy Removal
MO	Memory Overhead
NPI	Number of Pages Identified

CHAPTER I

INTRODUCTION

1.1 GENERAL

The Internet [1] is the collections of computer that are connected with each other and is used to provide information needs of users worldwide with the use of TCP/IP protocol suite. It provides access to documents that are hyperlinked together from the large information reservoir. This large information reservoir is known as World Wide Web or simply a web. WWW is a framework in which hyperlinked document are collected in the form of text, image and video. Earlier the information found manually over the web owing to availability of the small amount of information. Now the information available on the web is changing day by day due to the era of digital world which results in information overload. Hence, the number of users accessing the required information available on web is also increasing has become an issue. With this the use of the tools for information retrieval has rapidly increased. For accessing the information in easy and fast way various information tools are available like Meta search engine, web directories and search engine etc. The use of search engine [2] tool for information retrieval [4,5] is most popular these days.

1.2. SEARCH ENGINE

Search engine is tool used to find the information over the internet in an organized manner. Software used by search engine is known as crawler, which download the web pages by traversing the web. These web pages are than indexed in the search engine data base, which is build and maintained by crawler [3]. This is also called as spider or bot. A crawler takes list of URLs called seed URLs and crawls the web by downloading the corresponding wed pages in to the database. These web pages are further scanned for the URLs extraction from them. The same process of crawling is again carried out by adding these URLs to the seed URLs. Then crawler forward these search pages to the indexer. The indexer is a computer program which is used for indexing the web pages and it recognize the text, links together with other content in the page by storing them in to search engine database file. Therefore, the query may fetch by keyword search and if the search matches with contents then the corresponding page retrieved from the database.

The general architecture of search engine is shown in figure 1.1.

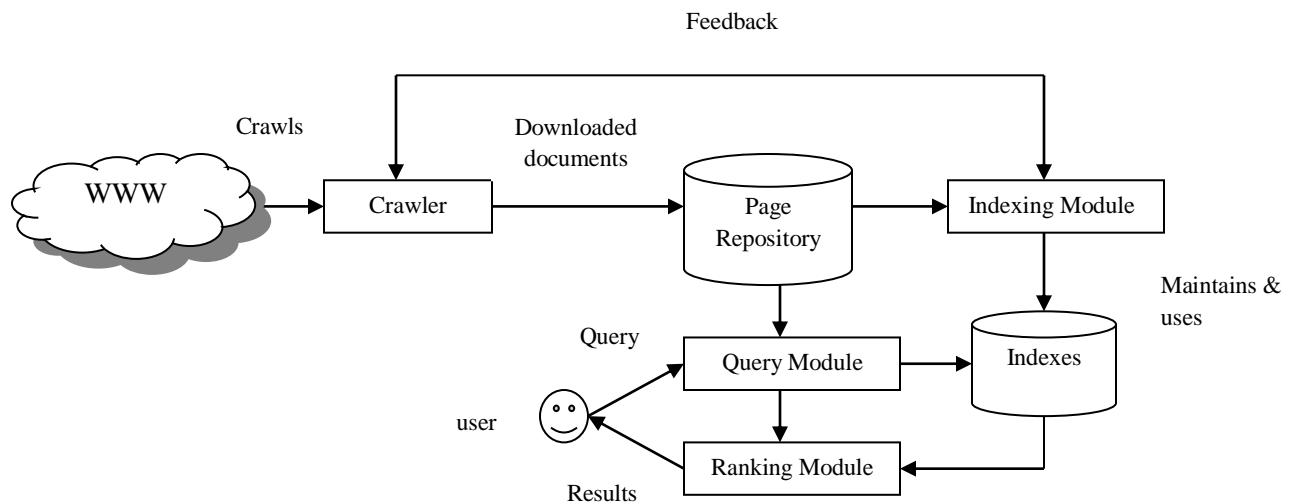


Figure 1.1: Components of Search Engine

The search engine is used to serve the users information need. The user fired query in terms of keyword on the search engine interface [6]. The query contains a single word or a sequence of words based on the information needs of the user. The query words may include white space and can be separated by AND, OR etc. Afterword the results provided to the user in the form of documents based upon the users request. The list of these documents are ordered or ranked according to their relevance. The most relevant documents are the topmost results in the retrieved list.

Search engines such as Google, Bing and Yahoo are the examples

1.3. MOTIVATION

Since, it is clear from the working of the search engine that it provides the user with a list of documents according to the query. The list of such documents may contain relevant and redundant documents. On the search interface if the user fires a query as a result, firstly the list of relevant documents is produced to the user's request. The list provides the information concern with the topic of interest. Since the retrieved list of documents thus produced is large in length and the user is concerned only in the relevant and novel information. Therefore, the user needs to

go through the whole list to extract the novel information. This results in the wastage of time and needs more effort by user. Thus for the solution to this problem there is need for a search system that provides satisfying information need in response to the user's query. The size of web is also increasing day by day with dynamic nature of documents results in following challenges in the design of efficient and effective search system.

1. *Crawls the quality pages:* In traversing the WWW (World Wide Web) [7] the crawler should obey a clear path. Hence, the quality of search engine may be maintained by diverting this towards downloading of the best pages first.

2. *Minimizing the overlapping of documents:* Multiple crawlers download the same page many times because the same page is available at multiple web sites. This downloading of the same web page is regarded as overlap problem, which results in redundant document retrieval. Hence, needs to build a search system to overcome this overlapping of documents such that duplicity in copies of same page multiple times may be avoided.

3. *Retrieving relevant and novel results:* The top results obtain by search system not always fulfill the need of user's request. In the retrieved pages there is no consideration of user preference and feedback for webpage ranking. Due to this large list is manually scanned by the user for selecting the needed information. It results in increasing the scanning time for results, which reduces the overall search experience of the user. Hence, need to improve the quality of return results such that in less number of clicks the information made available to user.

1.4 NOVELTY DETECTION TO IMPROVE THE SEARCH EFFICIENCY

The traditional search engine takes a query on the search interface and produces the list of relevant documents, which satisfied the user's information need. Now at this point one thing is clear that the list produced by the search engine may contains redundant documents. A redundant document is one which is relevant to user's information need but contains the same information as the previous document. Thus there is a need to make a search system which removed this already seen information by the user. Because this redundant information or documents take lot of user's time and effort for providing the required information need. Therefore the need for novelty detection [10] to improve the search efficiency in text documents is recommended.

Novelty detection [11] is the process to extract relevant and novel information from the result provided by search engine. It takes the ranked list provided by the tradition document retrieval system and further extracts the novel information to satisfy the user's request. Novel information is relevant to user's query and also provides new things as compared with the previous ones. Novelty detection is an important research area now a day due to the information overload over the web. With the novelty detection the user got new information with less effort and time. The process of novelty detection involves various novelties metric such as cosine similarity [13] by which a novelty score is calculated. This novelty score is than compared with a predefined threshold if the novelty score is above the threshold than the document is regarded as novel otherwise not novel.

Novelty Detection can be done at three levels

- **Event Level:** At the event level [12] the document is relevant and novel to topic or query but also related to new event. In other words new event present new information about an old event. The work at the event level novelty detection comes from the research Topic Detection and Tracking (TDT) [14].
- **Document Level:** In this novelty detection is performed at document level to classify the document is novel or not. The problem with this method is that if there are a vast number of documents collected in the system which also contains previously seen information; the incoming document has to compare with the entire document in the system. This is an unmanageable and complex task. To solve this problem novelty detection on sentence level has been done.
- **Sentence Level:** At this level sentence in a document is relevant and novel to the topic and also presented new information. The task is to extract relevant and novel sentences from the list of relevant documents with the given user's query. A novel sentence should be relevant to a topic and provides new information. The methods available to developed novelty detection at the sentence level [15, 16], the new word mostly contribute a lot to rank the sentence as novel.

1.5 OBJECTIVES OF THE WORK

The objective of this research is to design the efficient novelty detection techniques for the textual documents which help to perform the effective searching within less time and effort. To achieve this objective, the works on following goals have been performed.

- **To design and validate a technique for identifying the novel documents from the stream of documents:**

***Solution:** Novelty identification is accustomed to distinguishing novel information from an approaching stream of documents. In this proposed work, a novel methodology for document level novelty identification is discussed and proposed. This work first splits a document into sentences, decides the novelty of every sentence [26], then calculates the document level novelty score in view of an altered limit. Exploratory results on an arrangement of document demonstrate that this methodology beats standard document level novelty discovery as far as repetition exactness and excess review. This work applies on the document level information from an arrangement of documents. It is valuable in identifying novel data in information with a high rate of new documents. It has been effectively incorporated in a true novelty identification framework in the zone of information retrieval [8, 9].*

- **To design a novelty detection technique based on the extractive summary of the documents:**

***Solution:** This work provided the relevant and novel results to the user query using extractive text summarization [105,107]. Automatic summarization is one of the most attractive, and interesting topics for researchers to find out the relevant text from the huge bunch of documents. The key aim of this automatic text summarization is to provide the users to get their data in the minimum period of time. Novelty Detection [10] is a much needed requirement for good classification system. In this approach, the merging of two techniques extractive text summarization and general novelty detection is used. The proposed idea provides enhanced results when compared with already existing work.*

- **To design and validate a technique based on clustering of the documents to identify the novel documents:**

***Solution:** In this work, a clustering [89] based approach for novelty detection which will provide the relevant and novel information to the user query has been proposed. Firstly the incoming stream of documents based on user query related to a domain is clustered using k-means algorithm [83]. User makes a query based on specific domain using search engine and the first thirty retrieved results scraped out and store in a file on the disk. These documents are used to make ten clusters each containing of similar documents. Based on these clusters one cluster head is selected from each cluster which will provide ten documents. All of these ten documents having large distance as compared to with each other. This provided a list of ten novel documents based on the query. The proposed idea provides enhanced results when compared with the results of Bing search engine.*

- **To design and validate a novelty detection technique to identify the novel documents based on semantic similarity and text summarization using ontology**

***Solution:** Current web crawlers search the queries at very high speed but the problem of novelty detection or redundant information still persists. In this method, a new novelty detection mechanism is proposed which can be appended with current crawler. The proposed mechanism first summarizes the text based on ontology [72] then hash value is calculated using winnowing algorithm [77]. This hash value of document is matched with others using dice coefficient in order to calculate the similarity index. Based on the threshold chosen for similarity, the document is treated as novel or not. The proposed web crawler is implemented using SQL as backend and Visual Studio-2012 as frontend. The result shows that the proposed strategy reduces memory consumption and at the same time reduces the number of documents hence minimize the user time for searching the data from the results obtained. In addition, the proposed approach can be used with other search engines like Google, Yahoo, Bing and Alta Vista with the aim to minimize the redundant documents.*

1.6. ORGANIZATION OF THESIS

The thesis has been organized in the following chapters:

Chapter I: This chapter starts with brief introduction of Information Retrieval System. The Traditional search engine architecture and the need for novelty detection to refine the results of basic search system have been covered. The basic concepts of Novelty Detection in text documents have been discussed in this chapter.

Chapter II: This chapter contains the Literature Survey done on Novelty Detection in text documents. A detailed review of work done in this area has been provided. Different methods for novelty detection are also provided in a Table. The concept of document to sentence level novelty detection is introduced here.

Chapter III: In this Chapter a novel architecture of Document to Sentence level novelty detection in text documents has been designed and proposed. The architecture depicts different functional modules that are proposed and discussed. The results also have been provided with the proposed approach.

Chapter IV: This Chapter proposes a new architecture of Extractive text summarization based novelty detection in text documents. The details of various modules of this architecture and basic flow with algorithms have been discussed. The comparison of the results with the existing approach also been discussed.

Chapter V: This Chapter proposes the architecture of clustering based novelty detection in text documents. The details of this architecture and algorithms with basic components have been discussed. The comparison of the results with Bing search engine is also provided.

Chapter VI: In this Chapter the Semantic similarity and textual summary based novelty detection technique for textual documents is proposed. This technique is not based on simple

keyword indexing but uses the concept and context of words for indexing, providing efficient and fast access to the user. To provide user with a list of relevant and novel pages as a result of query entered by user is also provided. The Comparison of results with generic crawler and proposed crawler also been discussed.

Chapter VII: This Chapter concludes the outcome of the work proposed in this thesis. It also discusses the possibilities of future research work based on the proposed approaches.

CHAPTER II

LITERATURE SURVEY

2.1 INTERNET

The Internet [1] which is a huge arrangement of associated PC organizations. It utilizes TCP/IP Internet convention suite to worker millions and trillions of clients around the world. It is an assortment of some little, medium, enormous, private, public, government, scholastic and business networks which has an organization of organizations which are totally associated by huge scope of hardware, remote and optical systems administration advancements. The Internet gives foundation to help network activities like electronic mail and supports a gigantic assortment of hypertext archives called WWW [7].

The Internet had its foundations in 1960's when ARPANET (Advanced Research Projects Agency Network) [1], was made by the Pentagon's Advanced Research Projects Agency to give a safe and survivable interchanges networks for associations occupied with protection related examination. IP innovation was additionally evolved to make the organization worldwide. They indicated bundling of electronic messages, their tending to plans and how they are sent over the organization.

Since it was hard for clients to recollect complex IP addresses, Paul. V. Mockapetris [11] proposed a convention called DNS (Domain Name System Protocol). This convention furnished planning of IP addresses with space names. These area names were essentially mix of English characters. Since individuals effectively recall names rather than numbers. DNS turned out to be well known. Ultimately the National Science Foundation (NSF), set up a disseminated organization of organizations fit for dealing with far more noteworthy traffic. In 1985, NSFNET was made by NSF and from here the utilization of Internet for all instructive scholarly scientists, government offices, and global exploration associations began. Tim Berners Lee dealt with the possibility of Ted Nelson. The goal of his work was to permit colleagues of European lab to share their venture data. The venture created by Lee later established the framework for the advancement of World Wide Web. By the 1990's the Internet experienced touchy development. Web turned into the hotspot for offering better and cost free administrations. The figure 2.1 shows architecture for internet communication:

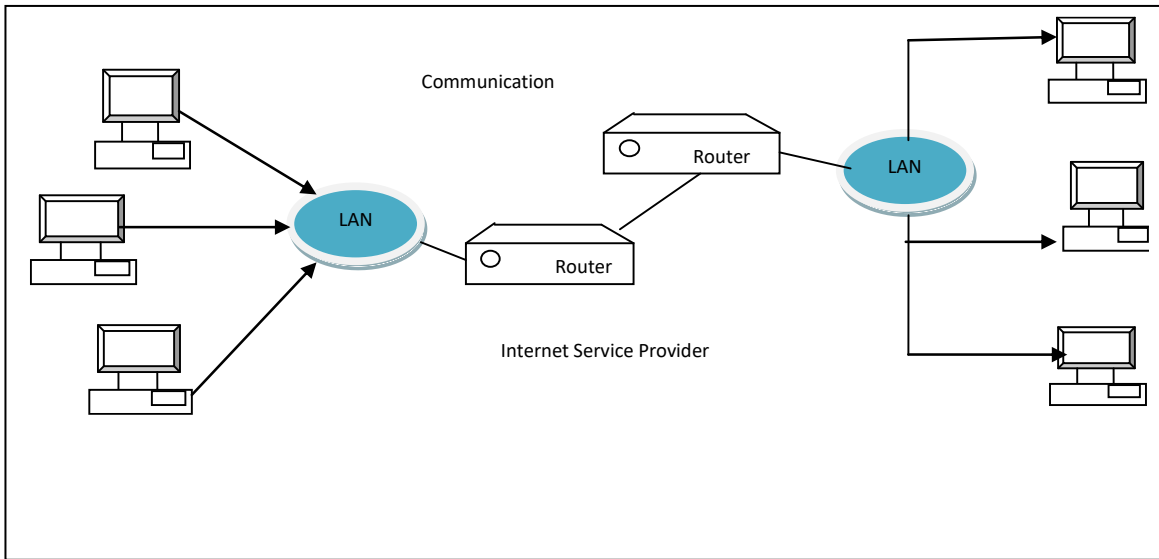


Figure 2.1: Architecture of Internet

In the current time, of Information Technology, Internet has made distances more limited and world more modest. It has become the amplest use hotspot for getting to a data. The Internet is presently being utilized in practically all circles going from training, government workplaces, design, correspondence, shopping, diversion and so forth. It has now become a key piece of human existence.

2.2 WORLD WIDE WEB

The WWW [7] is enormous assortment of hyperlinked documents that are made accessible with the assistance of Internet. Since Internet has the sites that contain website pages, consequently WWW is viewed as a piece of the Internet. HTML records containing labels for pictures, text, sound, video, and connection to different reports and so forth. Labels structure the foundation of the World Wide Web. Each website page has an interesting URL (it contains both the location of the PC on which it is put away and name of the record that contains the page) to separate itself from other organization assets. Toward the starting WWW facilitated static content pages just yet now it has been upset to give dynamic data that incorporates video, sound, mixed media, illustrations and 3-D livelinesss. Path back during World War II, Memex was created by Bush, which was a blend of transmission TV and microfilm innovation. Anyway the framework was lumbering and complex that never came into creation. Later Nelson presented Xanadu project that pre-owned hypertext archives unexpectedly. Yet, this undertaking was additionally

extremely mind boggling and was rarely figured it out. The thoughts given by Memex just as Xanadu affected Tim Berner Lee, who later developed the WWW. The significant segments that help perusing of site pages through Internet are Web workers, Web programs and HTTP convention. The worker stores the hypertext web archives. These records re recovered with the assistance of Hypertext transport convention. These records are seen on a unique programming application called Web-Browser. These product applications help Internet clients in looking, getting to and perusing pages. The first historically speaking internet browser created was Mosaic. Afterward, Netscape Navigator turned into a pioneer internet browser. An assortment of internet browsers that are common these days are Internet Explorer, Google Chrome, Bing, Firefox and so forth. In 1988, U.S Department of trade framed ICANN (Internet Corporation for Assigned Name and Number) to privatize the activity and enrolment of area names.

Figure 2.2 shows the essential engineering of WWW (W3) given by Tim Berners–Lee et al [7].

The World Wide Web Consortium (W3C) was established in October 1994 to advance the improvement of World Wide Web innovation.

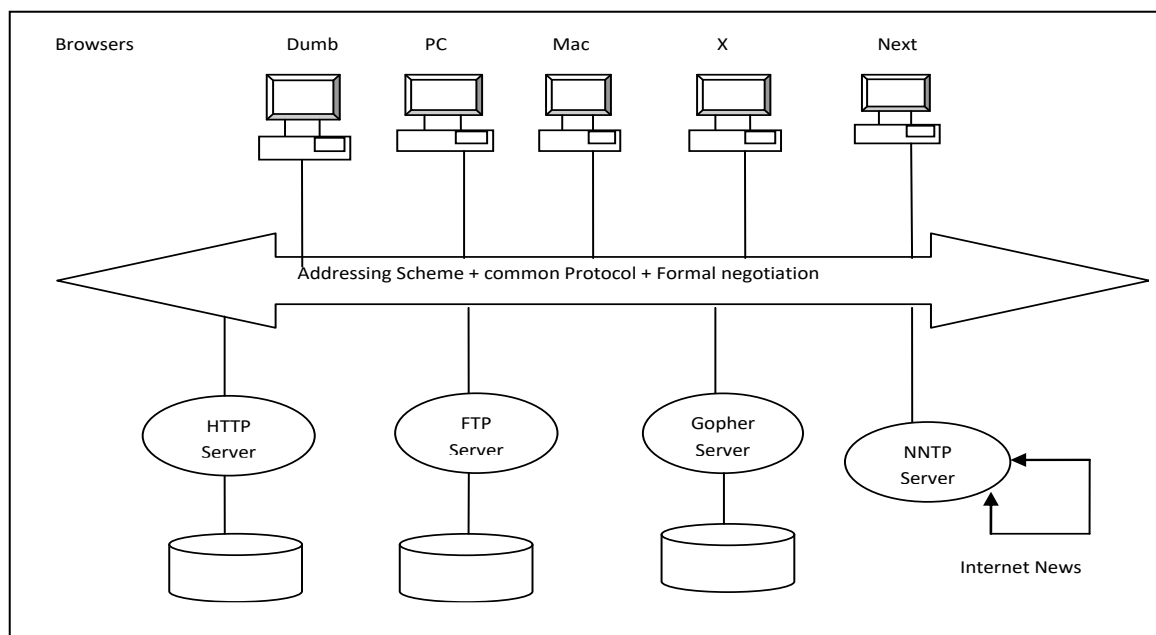


Figure 2.2: Basic Architecture of WWW

Since its inception in 1989, the size of Web has increase exponentially. More than half million web-sites exist on the Internet and each may contain numerous URL's. Google, the most popular search engine had 26 million web pages indexed in 1998 and now it indexes around 16 billion

web pages. It is thus very difficult for end users to access information from this huge information gateway. As a result there is a need for information retrieval tools which may help the end users to access the required information.

2.3 INFORMATION RETERIVAL

Information retrieval (IR) [9, 10] is the field of retrieving information from the different information sources available over the web in response to the user information need. IR takes a user's query and retrieves a list of documents satisfying the user's interest. The web is a large collection of web pages in the form of text, audio, imaged and video. The information available on the web may be in unstructured, structure and semi structure form. The unstructured form is not exactly clear thus can easily understand by the computer. The structure data is in the relational form maintained in the form of tables and clear. As the information size is increasing over the web it emerges the need of information retrieval.

The processing and filtering of the retrieved documents to fulfill the need of user's is performed by the information retrieval system. The task of information system starts by submitting the query on the search interface and the system retrieve the results to satisfying the information need. The grouping of the documents based on the contents of the documents is called clustering [89] and the process of classifying which document belongs to which class is called classification. The task of classification, initially start manually and then perform automatically. Upon submitting a query and the query match with the entity stored in the database. The query is formal way to satisfying the information need and the entity represents the objects in the data base in the form of text, image etc. The query matched with the several objects in the database not only uniquely matched with a object. Afterward a score is calculated by the information retrieval system of matched objects in the database and topmost documents reprieved by the system. The performance of the IR depends on the many factors that are discussed in the next section.

2.3.1 FACTORS AFFECTING PERFORMANCE IN IR

The performance of the IR depends on the many factors and some of them discuss here like precision and recall. These factors affected the performance in term of relevant and novel documents retrieval.

- **Precision**

Precision is the measure of number of relevant document retrieved by the system. It involves all the retrieved documents to calculate the fraction of relevant document retrieved by the system. For example, if the system having total 50 documents in the system and the 20 are the relevant documents. And the system returns only 10 documents as relevant document out of 50. This fraction of documents between relevant retrieved to the total number of documents is the precision of the system. It is defines as:

$$Precision = \frac{\text{Relevant documents retrieved}}{\text{Total number of retrieved documents}} \quad (1.1)$$

- **Recall**

Recall is the ratio of the total relevant documents in the system to the relevant documents retrieved. For example, if the system having total 50 documents in the system and the 20 are the relevant documents. Suppose the system returns only 10 documents as relevant to the user query, then fraction of 10 documents to the 20 documents is called recall. It is the measure of how many documents which are relevant to the query in the system has been successfully returns. It is defined as:

$$Recall = \frac{\text{Relevant documents retrieved}}{\text{Total number of relavent documents}} \quad (1.2)$$

Recall is also called the sensitivity. It is non-trivial to achieve 100 % recall but one must also consisering the non-relevant documents in to account. In the next section brief introduction about the tool for information reterival to be dicussed.

2.4 WEB DIRECTORIES

Web directories are used to search information over the web. These are organized in hierarchical tree structure based on the topic and subtopic of the web pages. They store the more general topic on the root of the tree and specific topic on the child of the root. If the user's want to search

the specific topic he must traversed the path from root to child node. Web directories are maintained manually; hence they cover the small portion of the web. Due to the small coverage web directories are not so popular these days for information retrieval.

2.5 META SEARCH ENGINE

Meta search is another tool for information retrieval over the web. Meta search engine find information simultaneously with the collaboration of other search engines. Such type of search engine does not maintain its own database but use the database maintained by search engine. Meta search engine is also not so preferable because they used high voluminous data from the diverse variety of search engines. Out of these information tools discussed above search engines are most preferable tool used for information retrieval now a days.

2.6 SEARCH ENGINE

Search engine is tool used to find the information over the internet in an organized manner. Software used by search engine is known as crawler, which download the web pages by traversing the web. These web pages are than indexed in the search engine data base, which is build and maintained by crawler. This is also called as spider or bot. A crawler takes list of URLs called seed URLs and craws the web by downloading the corresponding wed pages in to the database. These web pages are further scanned for the URLs extraction from them. The same process of crawling is again carried out by adding these URLs to the seed URLs. Then crawler forward these search pages to the indexer. The indexer is a computer program which is used for indexing the web pages and it recognize the text, links together with other content in the page by storing them in to search engine database file. Therefore, the query fired is processed by the query processor and based on keyword search if the search matches with contents then the corresponding page retrieved from the database.

2.7 NOVELTY DETECTION

As the size of web is increasing at higher rate due to the more and more information is uploading over the web. This information is flow between the user, services and clients via internet. There are different information resources available over the internet from where this information is access in the form of reports, articles and stories etc. by the different kind of users. Thus, in this

situation needs an automatic system based on novelty detection that provides updated information to the users. In the earlier days the information was accessed manually by maintaining the web directories etc. For example if in a library to maintain information about the variety of subjects a catalog was needed to categorize the material with different codes. In the era of internet the information is available in digital form; no need to use books or library for the required information need. Novelty detection [12] is a new field emerged from the last decade and an important topic for the researchers. Novelty means new answer in response to the question. The information retrieved by the system is regarded as novel if it contains new information as compared to the last information seen. Novelty is also measures the opposite of redundancy. The process of novelty detection start with a set of documents which are ranked based on their relevance score. A score is high for a document which is having more importance than the document having lower importance. Thus the task of novelty detection splits in two parts i.e. relevant document extraction and from that novel documents extraction. In the literature asymmetric measure and symmetric measure are there, which depends upon the ordering of the sentences in a document. A symmetric measure follows the ordering of the sentences. The cosine similarity [13] measures the similarities between the two vectors, if a sentence is represented by vector then similarity between the current sentence and previous sentence can be calculated. The main aim of novelty detection is to remove all the redundant and non relevant documents and presented novel documents to the user's based on their information need. In this way the user only read the needed documents and the redundant and non-relevant information is gone away. Novelty detection methods characterize as indicated by the accompanying general classifications, for example, Classification based, Neural network based, Support Vector Machine (SVM), Nearest Neighbor, Clustering and Statistical based strategy. These methods differ in way, how they find the novelty of the documents. The system can obtain relevant documents regarding given question and separate the non-relevant documents in the classification stage.

In the table no. 2.1 a brief overview of novelty detection methods are presented

Table 2.1: Methods for Novelty Detection

S.No.	Novelty detection Methods	Description
1.	Classification based	<ul style="list-style-type: none"> • Training of normal data is used to model this and new pattern is classified further. • The classifier is trained to check that the data is part of classifier or the new data.
2.	Neural network based	<ul style="list-style-type: none"> • In this method by using the normal data a neural network is trained than the output is checked according to the input. • A normal pattern is identified if the neural network accepts the input otherwise this is a novel pattern.
3.	Support vector Machine	<ul style="list-style-type: none"> • In this method data belongs to the low density region regarded as novel data and the data lies on high density area is the normal data. • In this method data belongs to the low density region regarded as novel data and the data lies on high density area is the normal data. • Minimum radius data is captured in a sphere. • To distinguish from spherical data a high dimensional space is used by non spherical data. • Accordingly, data comes within learned area is normal if not the data id novel
4.	Nearest Neighbour	<ul style="list-style-type: none"> • The data point that lies near to each other are the normal points and the points which fall away from each other are the novel points. • Distance based: In this a threshold value is used which signifies about the novelty of data points. The Distance between two neighbouring points above the threshold is novel otherwise not. • Density-based method: Data a point fall in low density area is the novel data otherwise not.
5.	Clustering	<ul style="list-style-type: none"> • In this technique, every ordinary data or information has a place with a group while the novel information doesn't have a place with any of the group. The groups or clusters are by and large huge. For each cluster, there exists a level of enrolment for every information point. A correlation of the level of enrolment is finished with the limit a value to discover cluster participation. In the event that an information point doesn't have a place with any of group, at that point it is novel.

6.	Statistical	<ul style="list-style-type: none"> • The Stochastic Distribution is utilized for recognizing the information. The Stochastic Distribution is a succession of random variables where variable accepts the value as 0 or 1. This is essentially utilized for modelling the framework which changes randomly.. Independently they are random however inside them, they denote a pattern. The information is formed on the standard of statistical properties. These properties are utilized for finishing up whether test information has a place with the same or distinctive distribution.
----	--------------------	--

In the following section a brief overview of novelty detection research on event level and sentence level is presented.

2.7.1 Event Level Novelty Detection

At the event level [15, 18] the document is relevant and novel to topic or query but also related to new event. In other words new event presents new information about an old event. The work at the event level novelty detection comes from the research Topic Detection and Tracking (TDT) [19]. This work is concerned with the online event/new story detection or new story and these events depend on the clustering method. Various models like vector space model and language model has been proposed to represent incoming stream of new documents/stories. These documents or new stories are grouped in to clusters. A new incoming story [20-27] is compared with the predetermine novelty threshold if the similarity score of the story is smaller than the threshold than it assigned in to closest cluster. And if this similarity threshold larger than the predetermine threshold than the story starts the new cluster and is regarded as new story.

In the work of Allan, Gupta, and Khandelwal [14, 16] the stories are presents each in turn and before the following story can be viewed as the sentences are score first. The assignment is to remove a single sentence inside a new topic for every occasion dependent on the temporal rundown.

2.7.2 Document Level Novelty Detection

In this novelty detection is performed at document level to classify the document is novel or not. The problem with this method is that if there are a vast number of documents collected in the system which also contains previously seen information; the incoming document has to compare

with the entire document in the system. This is an unmanageable and complex task. To solve this problem novelty detection on sentence level has been done.

2.7.3 Sentence Level Novelty Detection

At this level sentence in a document is relevant and novel to the topic and also presented new information. The task is to extract relevant and novel sentences from the list of relevant documents with the given user's query. A novel sentence [17] should be relevant to a topic and provides new information. The methods available for similarity computation in information retrieval that can also used to calculate similarity in novelty detection. In the sentence if new word comes they contribute high rank score for novelty detection. A sentence with high similarity with the query it increases the relevance of the sentence and if this score is also high with all history sentences than its relevance decreases.

Afterword the new word count appearing in the sentence to compute the redundancy of the sentence. This new redundancy or novelty measure is called New Information degree (NID) and its value is calculated in two ways. Firstly, take the ration of IDF (Inverse document frequency) values of all the new words in a sentence to the IDF values of all the words in the sentence. And secondly the NID value is calculated by taking the fraction of bigram of all the new words in the sentence over the all the bigram for words in the sentence. This work helps in computing the novelty by new word counts

Tsai and Zhang [26] have proposed the document-to-sentence model for document-level novelty identification. In this technique document level novelty perform effectively by utilizing sentence level method. The results based on APWSJ [26] data show that this technique perform better in terms precision and recall for redundancies than cosine similarity.

In other studies which utilizes the named entity [28] extraction and part of speech (POS) tagging model of novelty scoring used to find the novelty score of each sentence. Two different types of matrices are used i.e. unique comparison and second one is the important value. The unique comparison calculates the number of words and important value calculates the number of matched entities. Then the total match of words is calculated between the history document and tested document.

In the work of Soboruff and Harmarn [23] discussed the novelty track (TREC) [21-25] to introduce the concept of novelty detection. In this, after experimentation with TREC data [34-38] the novel sentences were extracted from the corpus based on query.

Li and Croft [46, 47, 48] have proposed in their work new definition for novelty identification. The query or question fire by user in response of this answer is presented is called novelty. They also suggested information pattern based novelty detection i.e opinion pattern or named entity. After experimentation they found that this method has enhanced the novelty for general topics.

In the work of Strokes et al [51] that used a lexical chain and text vector. The proposed method combined the lexical chain with the free text vector. The chain which is semantically related words in a text is called a lexical chain i.e. truck is lexical chain consist of vehicle, wheel, engine and automobile etc. In this chain each word is semantically related directly and indirectly. For example world net can be used to create the lexical chain and adopting this was result in marginal increase in effectiveness

Zhang et al [32] has proposed a method for novelty identification which was based on overlap-based redundancy calculation. In this method number of matching terms normalized by the length of the sentence was used to calculate the redundancy score between sentences. The threshold value used was 0.55 and the sentences with redundancy above the threshold were regarded as redundant. Thus these sentences eliminated from the repository

Tsai, Hsu and Chen [26] proposed the work in which cosine similarity function was used for detecting the novel sentences. In this work sentence are represented as vectors than the similarity score between the tested sentence and history sentence was introduced for novel sentence extraction.

Litkowski [37] has proposed the work on discourse entities in a sentence. Discourse entities are semantic objects. These objects have many syntactic relations in text document. For example a person name, a noun and pronoun all hails from single entity and considered as same discourse

entity. If new discourse entity found in a sentence with respect to all the history count of discourse entities than the sentence is regarded as novel.

In the work Eichmann et al [38] has introduced in the sentence the concept of named entities and noun phrases. The task of novel sentence detection was based on the he count of the new named entities and noun phrases. If they are greater the preset count than the sentence was regarded as novel otherwise not.

Allan et al work [40] on temporal summaries of new topics, the task is to extract a single sentence from each event within a news topic, where the stories are presented one at a time and sentences from a story must be ranked before the next story can be considered. Novelty is an important characteristic of sentence selection. They proposed two approaches for measuring novelty. The first approach estimates the probability that two sentences discuss the same event by the probability that the later sentence could arise from the same language model as the earlier sentence. The second approach is the same as the first approach except that the sentence is compared to clusters and there is more information in a cluster to estimate probabilities. The second measure gives better performance than the first approach, which indicates the clustering is useful for modeling the events.

In the work of Zhang et al [39] the removal of redundant documents has been discussed on five redundancies measures i.e. three KL-divergence based measures with three language modeling variations. The other two redundancies measures were cosine similarity and set difference measure. The KL-divergence involves Dirichlet smoothing, shrinkage smoothing and three-component mixture language model. It was found after the experimentation that the redundancy measure based on mixed language model and cosine similarity measures provides best results in detection of redundant documents.

Blended metrics were proposed in [43], whose methods consolidate both cosine similarity and new word check measurements for novelty mining. The main features of the blended metrics are combining the measurements from multiple metrics similar to constructing a classifier using multiple features rather than only one feature. Since the qualities of the two measurements

contrast, a few examples that cover in the one-dimensional feature space could be detachable in the two-dimensional element space. The preferred position is that the choice limit in the two-dimensional element space which is characterized by the mixing technique and edge can be more adaptable than that in the one-dimensional component space [43].

Fernandez and Losada [42] study have addressed Local Context Analysis (LCA) to detect new and relevant sentences within documents related to a certain topic. LCA is beneficial to researchers in different areas of study, such as text summarization, information retrieval, Web search engines, question answering systems, etc. The core idea of this method is based on a common term from the top-ranked relevant documents that tend to co-occur within query terms within the top-ranked documents.

Li and Craft 2005, Gabrilovich 2004, Kwee 2009 and Bentivogli 2011 [49, 56, 58, 62]: have done the work on removal of duplicate documents for novelty and redundancy identification in adaptive filtering. In the novelty sub topic of RTE- 6 and 7 was explored in textual based sentence level novelty mining.

Saed Alqaraleh and Omar Ramadanin [63, 64]: The work has been done with the multimedia files like images, music, and videos. The objective was to improve the efficiency of multimedia search engines by eliminating repeated occurrences.

In the work of Sravanthi, P., & Srinivasu, B.[65]: has introduced the method for word co-occurrence. In this method the word order was ignored for the sentence. Thus in the context of sentence it does not count the meaning of the word.

Dasgupta and Dey, 2016 [66]: has been carried out the work on the problem of novelty detection at the document level. In the work of Tirthankar Ghosal, amrita Salam , Swati Tiwari , Asif Ekbal, Pushpak Bhattacharyya, 2018 [67]: they have suggested the problem of novelty by building the resource through event specific crawling of new document. The work was on document level novelty detection from many domains in a periodic way. TAP DND 1.0 corpus was used by incorporating the supervise machine learning method.

The other related contribution of work done in the area of Duplicate Detection by the following authors as:

Some documents or web pages are the mirror image of the others. Whereas some documents are not exact copy of other document but differ in some part are called near duplicate. These more numbers of duplicates pages require more memory space and effect the speed together with cost in providing the results. Thus recognizing these duplicate pages save cost and memory space. These pages are generated due mirror versioned or imitation of documents. The quality of results produced by search engines can be enhanced after the removal of these redundant pages. For removing such duplicate pages various methods provided by many authors:

In the work of Broder et al [68] shingles were produced with the sequence of all nearby words. The method has been proposed which tells the documents with same shingles are treated as similar, whereas the overlap shingles are treated as exactly same. This method works better with smaller length documents

Theobald et al. Spotsigs [69] have proposed the duplication detection method. In this method spot signatures are used by removing the stop word from anchors. After an anchor excluding stop words, than by using k token right these spot signature are created. Hashing is done on these signatures to reduce the signature vectors length. Afterwords the generated hash codes are compared if the difference between them is smaller than the preset threshold, these pages as counted as near duplicate. Hence these web pages are eliminated from the depositary.

In the work of Fetterly et al [70] for the identification of near duplicate pages has used super shingling method. The work proposed by Hannaneh Hajishirzi et.al [71] for specific domain to identify of near duplicates. Hence for similarity computation vectors mapped to smaller hash values to increase the efficiency in detection

2.8 SUMMARY

It is apparent from the literature survey the web crawler process the query at exceptionally fast yet the issue of novelty detection actually continues. Anyway till date no web crawler has had the option to furnish exact and precise outcomes concerning the query. The users actually need to check the all return results by refining them physically to get the needed information. Thus,

Design of Novelty Detection Techniques for Optimized Search Engine results is being proposed to dispose of the downloading of excess documents and giving the novel results.

In the figure 2.3 the complete flow of work is shown.

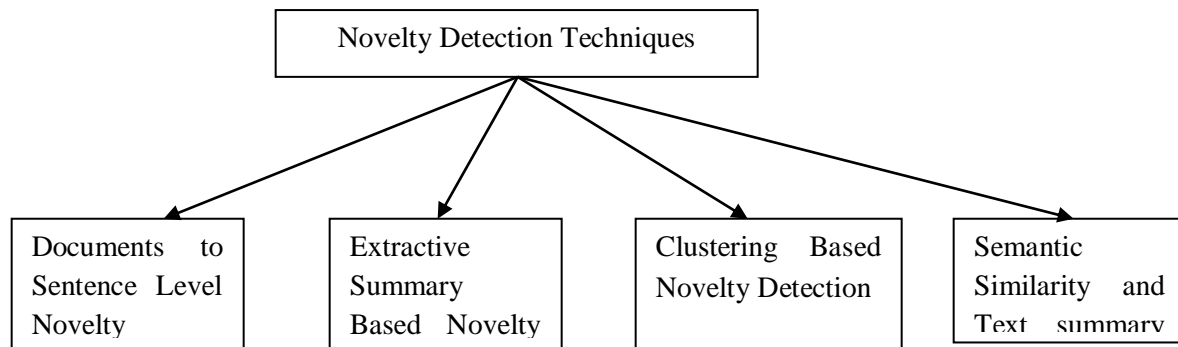


Figure 2.3: Complete Work Flow

The following are the techniques that are proposed to provide the relevant and novel text document retrieval:

- Document to Sentence level Novelty Detection Technique
- Text Summarization based Novelty detection
- Clustering based Novelty detection in Text Documents
- Semantic Similarity and Text Summarization based Novelty Detection

In the coming next chapters the architecture of novelty detection techniques with detailed description is provided.

CHAPTER III

DOCUMENT-TO-SENTANCE LEVEL NOVELTY DETECTION

3.1 INTRODUCTION

In the previous chapter, the history and evolution of the novelty detection methods at the two levels i.e. sentence level and event level has been discussed. Novelty detection is accustomed to distinguish novel data from an approaching stream of documents. In this chapter a novel methodology for document level novelty identification by utilizing Document to-sentence level strategy has been proposed. The work first splits a document into sentences [17], decides the novelty of every sentence, then calculates the document level novelty score. Exploratory results on an arrangement of document demonstrate that the proposed methodology beats standard document level novelty discovery as far as repetition exactness and excess review. The work has been applied on the document level information from an arrangement of documents. It is valuable in identifying novel data in information with a high rate of new documents.

3.2 PROPOSED DOCUMENT TO SENTENCE LEVEL NOVELTY DETECTION

One of the most important points of concern in the proposed work is to how the idea of novelty detection refines existing search- engine results. A novel approach to detect the novelty in the documents is presented in this chapter. Many important applications have used novelty detection in order to reduce redundant and non-relevant information presented to users of the document retrieval systems. In this study document level novelty detection has been proposed and the algorithm is aimed at removing the redundancy of the results and increasing the speed to find the novel information in the documents. To increase the speed, novelty score of documents is calculated in the proposed algorithm by using the sentence segmentation instead of using whole document. Sentence segmentation [17] has been done by pre-processing the document and then splitting the document into sentences.

3.3 NOVELTY DETECTION ALGORITHM

Document to Sentence Level Novelty Detection algorithm is a proposed detection algorithm which is used to find whether a document gives novel information or not when compared with

the history documents. The algorithm first splits the document into sentences, determines the novelty score of each document based on a fixed threshold.

For similarity computation [13] sentences are then compared with all the history sentences. To compute the nature of document, similarity is converted to novelty score for each sentence. A minimum value is selected out of the novelty values and finally the decision has to be made.

3.3.1 Architecture of Proposed Novelty Detection Method

The architecture for the proposed system is shown in figure 3.1. In this system user enters a query in the form of a new document; this is passed to segmentation module.

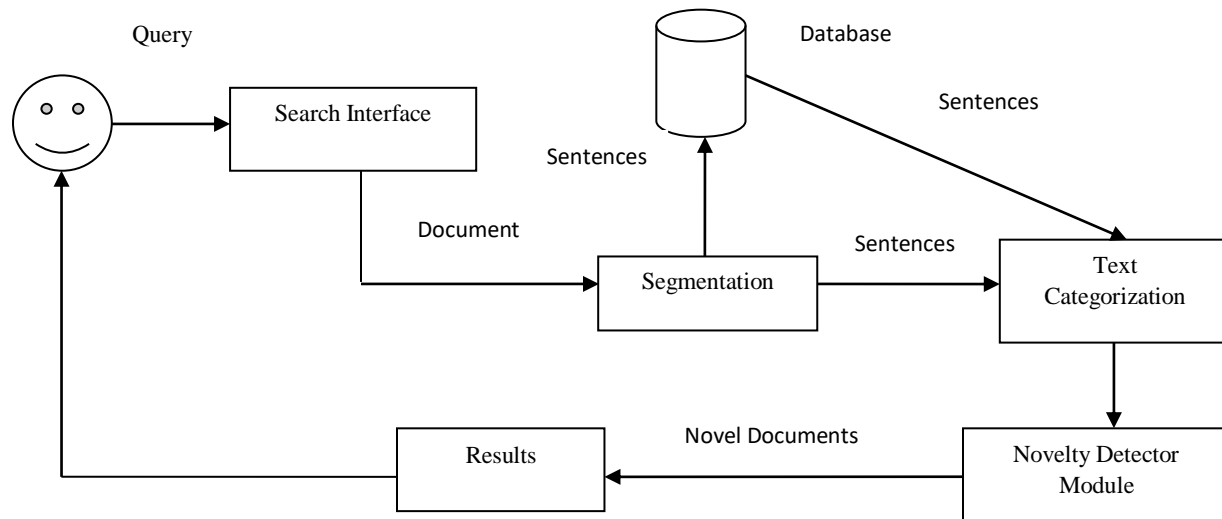


Figure 3.1: Architecture of Proposed System

This module breaks the documents into sentences and stored them in the database. Now the sentences are passed through text categorization module where stemming and stop word removal processing is done. The detector module decides the novelty of the document. At last the result is passed to the user by result module. Now, various components and novelty process are explained below in detail to have an understanding of the proposed detection system:

3.3.1.1 Segmentation Module

This module breaks the documents into sentences and these sentences are stored in the database for further processing. For example the document “Karnal is a city located at the centre of the Haryana. The population of Karnal is approximately 2.87 lakhs” is break in to sentences like:

Karnal is a city located at the centre of the Haryana.

The population of Karnal is approximately 2.87 lakhs.

3.3.1.2 Text Categorization

This module helps in pre-processing the sentences. It removes the stop words from the sentences before they move to the next module. For example ‘the pen is blue’ so after removal of stop word can and be it results in pen, blue.

3.3.1.3 Novelty Detector Module

This module helps in finding the novelty of the document. The process of this module is as shown in figure 3.2. The document is segmented into sentences; compute the novelty score of each sentence by using the sentence-level novelty detection module. Then, the average of novelty score is compared with a fixed threshold value, if the value of novelty score is greater than the threshold value then the document is considered as novel otherwise not.

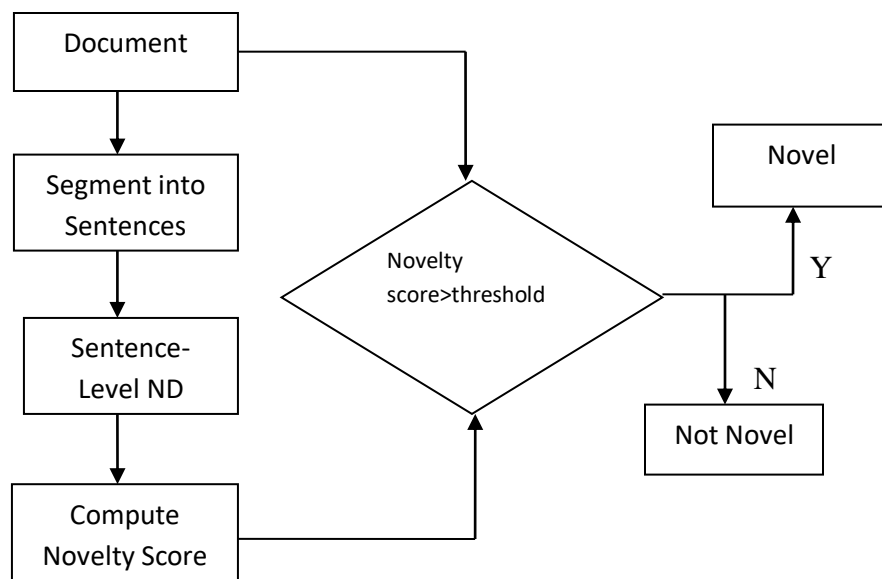


Figure 3.2: Novelty Detection Process

For measuring the geometric distance cosine similarity [13] is used between the sentences. From previous studies, it has been cleared that the cosine distance metric showed a good performance

on both document and sentence-level novelty detection. Hence, in this work cosine similarity is used to predict whether an incoming sentence contains enough novel information compared to a set of history sentences.

Cosine similarity is a symmetric measure related to the angle between two vectors. If we represent a sentence s as a vector $s = [x_1(s), x_2(s), \dots, x_n(s)]^T$, then the cosine similarity between two sentences is defined as:

$$\cos(st, si) = \frac{\sum x_k(st) x_k(si)}{||st|| \cdot ||si||} \quad 3.1$$

Where, the value of k lies from 1 to n . Cosine Similarity measures of cosine of angle between two vectors and the values is in the range of $(-1, 1)$. Two sentences have considered for the cosine value calculation as below:

S_1 . Ram goes to marry Panjabi girl, a girl.

S_2 . Shyam goes to Punjab to find Ram.

Calculating the terms and their corresponding frequencies for s_1 and s_2 :

TABLE 3.1: Term and Frequencies of Sentence ' s_1 ' for Calculating Cosine Similarity

TERMS	Ram	Goes	to	marry	Punjabi	girl	a
FREQUENCIES	1	1	1	1	1	2	1

Also, calculate the term and frequencies for the sentence ' s_2 '

TABLE 3.2: Term and Frequencies of Sentence ' s_2 ' for Calculating Cosine Similarity

TERMS	Shyam	goes	to	Punjab	Find	Ram
FREQUENCIES	1	1	2	1	1	1

From the table 3.1 and 3.2 the total number of terms in sentence ' s_1 ' is 8 and in ' s_2 ' is 7.

TABLE 3.3 Term and Frequencies of Sentence ‘s1’ and ‘s2’ together

TERMS	Ram	goes	To	marry	Punjabi	girl	Shyam	Punjab	Find
Frequency in S1	1	1	1	1	1	2	0	0	0
Frequency in S2	1	1	2	0	0	0	1	1	1

Now, the Cosine product is calculated between the sentences s1 and s2

The vectors for sentence s1 and sentence s2 can be represented as:

$C = [1,1,1,1,1,2,0,0,0]$, $D = [1,1,2,0,0,0,1,1,1]$.

Thus, *COS* of angle between two vectors is calculated using the *eqn. 3.2*.

$$\text{Similarity} = \cos(\theta) = \frac{C \cdot D}{||C|| ||D||} = \frac{\sum_{i=1}^n C_i \times D_i}{\sqrt{\sum_{i=1}^n (C_i)^2} \times \sqrt{\sum_{i=1}^n (D_i)^2}} \quad (3.2)$$

$$\cos \theta = \frac{1 \times 1 + 1 \times 1 + 1 \times 2 + 1 \times 0 + 1 \times 0 + 2 \times 0 + 0 \times 1 + 0 \times 1 + 0 \times 1}{\sqrt{1^2 + 1^2 + 1^2 + 1^2 + 2^2} \times \sqrt{1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 1^2}} \quad (3.3)$$

Therefore, the value of *COS* θ is calculated as 0.377.

For the novelty detection task, in order to measure the degree of novelty directly, the cosine similarity score is converted in to the novelty score simply by one minus cosine similarity score. Cosine similarity metric compares the current sentence with each of its history sentences individually, where the minimum novelty score among them has been chosen as the novelty score of the current sentence.

$$\text{Novelty Score}(st) = \min[1 - \cos(st, si)] \quad (3.4)$$

Given below are three simple documents that occurred one after the other:

D₁: Karnal is a city located at the centre of the Haryana. The population of Karnal is approximately 2.87 lakhs.

D₂: It lies 150 kilometres from Delhi. Karnal is a city located at the centre of the Haryana.

D₃: Karnal lies 150 kilometres from Delhi. The population is approximately 2.87 lakhs.

The proposed algorithm is shown in figure 3.3 which predicted the novelty of the document, whether it is novel or not. The algorithm when applied on the documents D₁, D₂ and D₃ it splits the documents in to sentences then compare first sentence of a document to all the sentences of other documents and so on.

3.3.1.4 Algorithm for Document Level Novelty detection

Algorithm: (N=set of documents, newDoc)

```
1. Begin
2.   for( i ← 1 to N)
3.     Begin
4.       for(sen_doc ← 1 to sen_newdoc)
5.         Begin
6.           for( sen ← 1 to sen_idoc)
7.             Begin
8.               Cos_sim[] ← find cosineSimilarity(sen, sen_doc)
9.             End
10.            maxCos[] ← find maximum value from cos_sim[]
11.            noveltyScore[] ← min[1- maxCos[]]
12.          End
13.        End
14.    avgNovel ←  $\sum \text{noveltyScore[]} / N$ 
15.    If(avgNovel > Threshold)
16.      Return newDoc is novel
17.    Else
18.      Return newDoc is not Novel
19. End
```

Figure 3.3 Algorithm for Proposed Work

When the general novelty detection method is directly applied, the document D₃ is quite natural to be predicted as a novel because it contains new information in comparison to D₁ and D₂

individually. But if these documents are segmented into sentences, D_3 will be correctly predicted as redundant because all its sentences have appeared in the previous sentences.

3.4 SIMULATION OF THE PROPOSED APPROACH

In this the proposed scheme is implemented using various examples. This section shows the performance enhancement of this approach.

3.4.1 Simulation

In this chapter the proposed algorithm is simulated and implemented by using two examples. User choose a new document and that document is compared with three documents. The result is computed by finding the novelty score for each document based on a fixed threshold.

3.4.2 Example 1

Step 1- 3 documents (N1, N2, N3) are taken for the basic analysis.

Our society is suffering from various social evils at the moment. The dowry system is common almost in all parts of India. Dowry has been stated as "the value paid by the parents for getting their daughters the place of a daughter-in-law". Parents pay huge sums of money so that their daughters may secure a satisfactory and permanent post. The groom's parents try to mine the maximum from a matrimonial association. They insist on getting huge amount of price, luxury items like television sets, VCR's, refrigerators, cars, and even houses.

Figure 3.4 Document N1

Originally parents of the bride would give their daughters present, ornaments, and other necessary articles of daily use with happiness. These few things were meant to contribute to a happy family life. The rich of our society gave this custom a design to fill pockets of the parents of the bridegroom. In due track of time demand for the dowry became most critical condition of the marriage settlement. The demand in cash which depends upon the merits of the boy and the status of the family become a terror for the society. Middle class people became main target of the attack. The devil of dowry has place an end to the pleasure of numerous couples even after marriage.

Figure 3.5 Document N2

Dowry has become a very common word and it is practiced in Indian society without any inhibitions or ill feelings. Dowry is a payment from the bride's family to the groom or groom's family at the time of marriage. Upon marriage, daughters are given all modern household gadgetry as dowry such as furniture, crockery, electrical appliances (in recent years refrigerators, television etc.) as well as personal items of clothing, jewellery and cash. Some parents also give a car among dowry items. The price of the dowry depends on the jobs the grooms may be holding at the time of marriage. When demands for dowry are not met, the bride is matter to pain, and often even killed. Most of these suicides are by hanging oneself, poisoning or by fire.

Figure 3.6 Document N3

Step 2- Now user selects a new document (newDoc)

Dowry has been defined as "the price paid by the parents for getting their daughters the post of a daughter-in-law". In due course of time demand for the dowry became most essential condition of the marriage settlement. The groom's parents try to extract the maximum from a matrimonial alliance. The amount of the dowry depends on the jobs the grooms may be holding at the time of marriage. The devil of dowry has put an end to the happiness of several couples even after marriage. When demands for dowry are not met, the bride is subject to torture, and often even killed.

Figure 3.7 New Document

Step 3- All the documents are segmented into sentences and each sentence of the new document are compared with all the sentences of history documents.

Our society is suffering from various social evils at the moment.

The dowry system is common almost in all parts of India.

Dowry has been stated as "the value paid by the parents for getting their daughters the place of a daughter-in-law".

Parents pay huge sums of money so that their daughters may secure a satisfactory and permanent post.

The groom's parents try to mine the maximum from a matrimonial association.

They insist on getting huge amount of price, luxury items like television sets, VCR's, refrigerators, cars, and even houses.

Figure 3.8: Sentences of Document N1

Originally parents of the bride would give their daughters present, ornaments, and other necessary articles of daily use with happiness.

These few things were meant to contribute to a happy family life.

The rich of our society gave this custom a design to fill pockets of the parents of the bridegroom.

In due track of time demand for the dowry became most critical condition of the marriage settlement.

The demand in cash which depends upon the merits of the boy and the status of the family become a terror for the society.

Middle class people became main target of the attack.

The devil of dowry has place an end to the pleasure of numerous couples even after marriage.

Figure 3.9: Sentences of Document N2

Dowry has become a very common word and it is practiced in Indian society without any inhibitions or ill feelings.

Dowry is a payment from the bride's family to the groom or groom's family at the time of marriage.

Upon marriage, daughters are given all modern household gadgetry as dowry such as furniture, crockery, electrical appliances (in recent years refrigerators, television etc.) as well as personal items of clothing, jewellery and cash.

Some parents also give a car among dowry items.

The price of the dowry depends on the jobs the grooms may be holding at the time of marriage.

When demands for dowry are not met, the bride is matter to pain, and often even killed.

Most of these suicides are by hanging oneself, poisoning or by fire.

Figure 3.10: Sentences of Document N3

Dowry has been defined as "the price paid by the parents for getting their daughters the post of a daughter-in-law".

In due course of time demand for the dowry became most essential condition of the marriage settlement.

The groom's parents try to extract the maximum from a matrimonial alliance.

The amount of the dowry depends on the jobs the grooms may be holding at the time of marriage.

The devil of dowry has put an end to the happiness of several couples even after marriage.

When demands for dowry are not met, the bride is subject to torture, and often even killed

Figure 3.11: Sentences of New Document

Steps 4- Sentences of N1 document are taken one by one.

Step 5- Find Cosine Similarity of each sentence

Table 3.4: Cosine Similarity Values of N1 Document

ND1	.842	.389
ND2	.428	.11
ND3	.85	-
ND4	.127	-
ND5	.427	.117
ND6	.252	.06

Table 3.5: Cosine Similarity Values of N2 Document

ND1	.334	-
ND2	.904	.476
ND3	.23	-
ND4	.589	-
ND5	.476	.857
ND6	.265	.265

Table 3.6: Cosine Similarity Values of N3 Document

ND1	.132	-
ND2	.159	-
ND3	0	-
ND4	.667	.975
ND5	.278	-
ND6	.882	-

Step 6- Now the maximum values of cosine similarity from each table is selected

Table 3.7: Maximum Cosine Values

	N1	N2	N3
ND1	.84	.33	.13
ND2	.43	.90	.16
ND3	.85	.23	0
ND4	.13	.59	.98
ND5	.43	.86	.28
ND6	.25	.26	.88

Step 7- Find novelty score for each document

Table 3.8: Novelty scores for N1, N1, N3

NEW DOCUMENT	N1	N2	N3
ND1	.16	.67	.87
ND2	.57	.10	.84
ND3	.15	.77	1
ND4	.87	.41	.02
ND5	.57	.14	.72
ND6	.75	.74	.12

Step 8- Now compute minimum novelty score for each document

Table 3.9: Minimum Novelty Scores

ND	N1	N2	N3
	.15	.10	.02

Step 9- Find the average novelty score

$$\text{avgNovel} = (0.15 + 0.10 + 0.02) / 3$$

$$= 0.27/3 = 0.09$$

Step 10- Now we compare the average novelty score with the fixed threshold value

Threshold = 0.45

avgNovel = 0.09 which is less than the threshold value

So, new document ND is not novel.

3.4.3 Example 2

Step 1- 3 documents (M1, M2, M3) are taken for basic analysis

Sport refers to a competitive physical activity. Sport is generally recognized as activities based in physical athleticism or physical dexterity. Sports are usually governed by rules to ensure good competition and consistent adjudication of winner. Records of performance are often kept and reported in sport news. Sport comes from Old French disport meaning leisure

Figure 3.12: Document M1

Health is the level of functional or metabolic efficiency of a living being. It is the general condition of person's mind and body, i.e. free from illness, injury. It is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity. Systematic activities to prevent health problems and promote good health in humans are undertaken by health care providers. Applications with regard to animal health are covered by veterinary sciences.

Figure 3.13: Document M2

Pollution is the introduction of contaminants into the natural environment that causes adverse change. Pollution can take the form of chemical substances or energy, like noise, heat or light. Pollution is often classed as point source or nonpoint source pollution. A pollutant is a waste material that pollutes air, water or soil. Three factors which determine the severity of a pollutant are: its chemical nature, the concentration and the persistence.

Figure 3.14: Document M3

Step 2- Now user selects a new document (newDoc)

Healthy human development is a necessary foundation for all development progress. Without healthy populations, the achievement of development objectives will be out of reach. Good health is fundamental to the ability of individuals to realize their full human potential. It is also a crucially important economic asset. Low levels of health impede people's ability to work and earn a living for themselves and their families.

Figure 3.15: New Document

Step 3- All the documents are segmented into sentences and each sentence of the new document are compared with all the sentences of history documents.

Sport refers to a competitive physical activity.

Sport is generally recognized as activities based in physical athleticism or physical dexterity.

Sports are usually governed by rules to ensure good competition and consistent adjudication of winner.

Records of performance are often kept and reported in sport news.

Sport comes from Old French *desport* meaning leisure

Figure 3.16: Sentences of Document M1

Health is the level of functional or metabolic efficiency of a living being.

It is the general condition of person's mind and body, i.e. free from illness, injury.

It is a state of complete physical, mental, and social well-being and not merely the absence of disease or infirmity.

Systematic activities to prevent health problems and promote good health in humans are undertaken by health care providers.

Applications with regard to animal health are covered by veterinary sciences

Figure 3.17: Sentences of Document M2

Pollution is the introduction of contaminants into the natural environment that causes adverse change.

Pollution can take the form of chemical substances or energy, like noise, heat or light.

Pollution is often classed as point source or nonpoint source pollution.

A pollutant is a waste material that pollutes air, water or soil.

Three factors which determine the severity of a pollutant are: its chemical nature, the concentration and the persistence.

Figure 3.18: Sentences of Document M3

Healthy human development is a necessary foundation for all development progress.

Without healthy populations, the achievement of development objectives will be out of reach.

Good health is fundamental to the ability of individuals to realize their full human potential.

It is also a crucially important economic asset.

Low levels of health impede people's ability to work and earn a living for themselves and their families.

Figure 3.19: Sentences of New Document

Steps 4- Sentences of m1 document are taken one by one.

Step 5- Find Cosine Similarity of each sentence

Table 3.10: Cosine Similarity Values of M1, M2, M3

New doc/Old Doc	M1	M2	M3
ND1	0	0	0.1
ND2	0	0.53	0.27
ND3	0.3	0.2	0.1
ND4	0	-	-
ND5	0	0.3	0.2

Step 6- Find novelty score for each document

Table 3.11: Values of Novelty Scores

New Doc/Old Doc	M1	M2	M3
ND1	1	1	0.9
ND2	1	0.47	0.73
ND3	0.7	0.8	0.9
ND4	1	-	-
ND5	1	0.7	0.8

Step 7- Now compute minimum novelty score for each document

Table 3.12: Minimum Novelty Scores

ND	M1	M2	M3
	0.7	0.47	0.73

Step 8- Find the average novelty score

$$\text{avgNovel} = (.7 + .47 + .73) / 3 = 0.63$$

Step 9- Now we compare the average novelty score with the fixed threshold value

$$\text{Threshold} = .45$$

avgNovel = .63 which is more than the threshold value. Hence, new document ND is novel.

From the examples the threshold value 0.45 is the standard value for better result. The table 3.12 shows about the selection of the threshold when applied on the three documents with the variation in threshold value

Table 3.13: Comparison of Threshold Values

Result with 0.35	Result with 0.45	Result with 0.55
Document 1:Result provided the 4 novel sentences	Result provided the 4 novel sentences	Result provided the 4 novel sentences
Document 2:Result provided the 2 novel sentences	Result provided the 2 novel sentences	Result provided the only 1 novel sentences
Document 3:Result provided the 2 novel sentences	Result provided the 2 novel sentences	Result provided only 1 novel sentence.
Document 4:Result provided the 3 novel sentences	Result provided the 2 novel sentences	Result provided the 2 novel sentences
Document 5:Result provided the 4 novel sentences	Result provided the 3 novel sentences	Result provided the 3 novel sentences

3.5 IMPLEMENTATION AND PERFORMANCE EVALUATION

For the experiment the query terms selected were ,*'Diwali'*,*'Holi'*,*'Cricket'*,and *'Pulwama attack'*. First 30 pages of search results were stored in repository for testing the proposed approach. The implementation includes Java, PHP, HTML, and Microsoft Access.

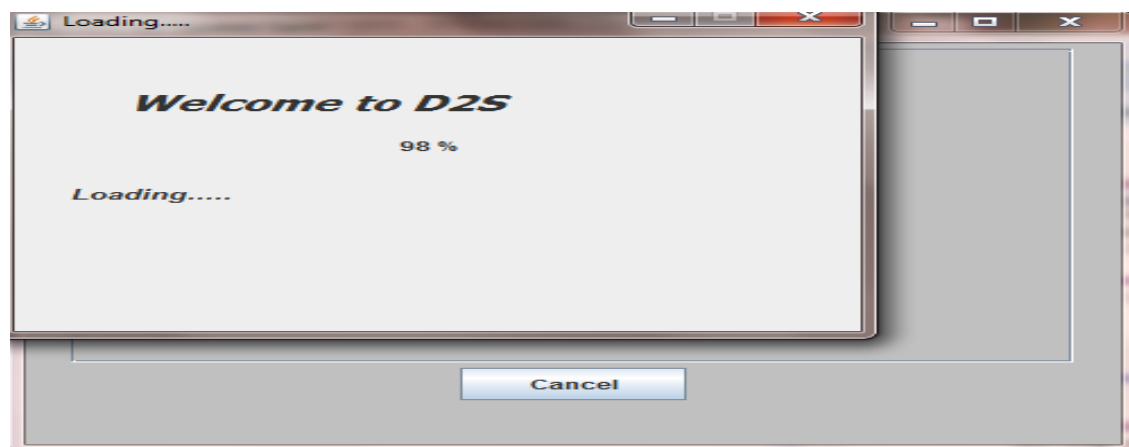


Figure 3.20: Query Interface

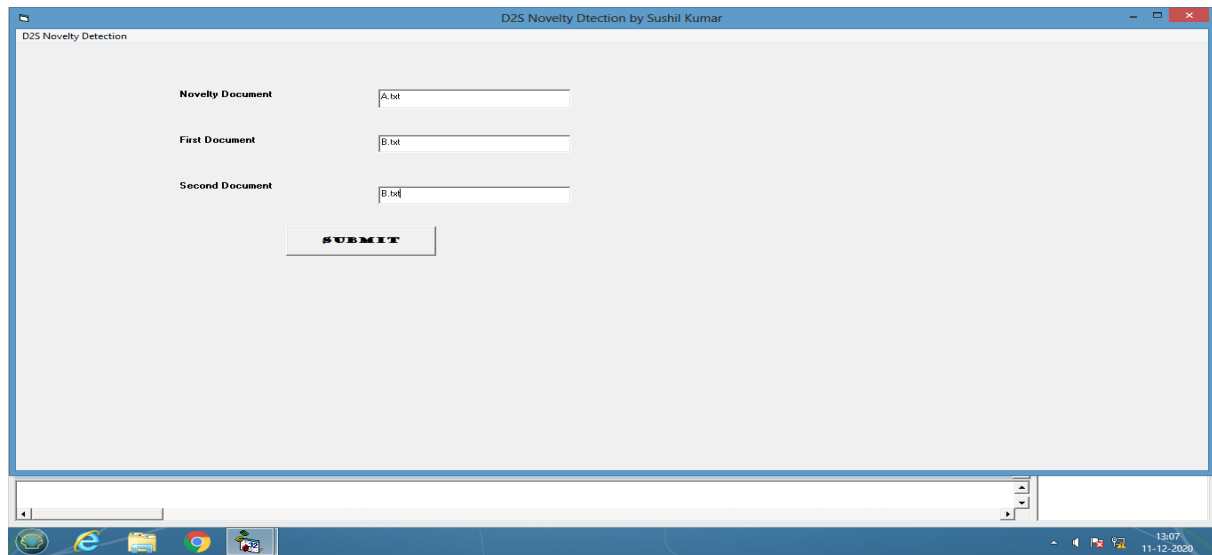


Figure 3.21: Input documents for Processing

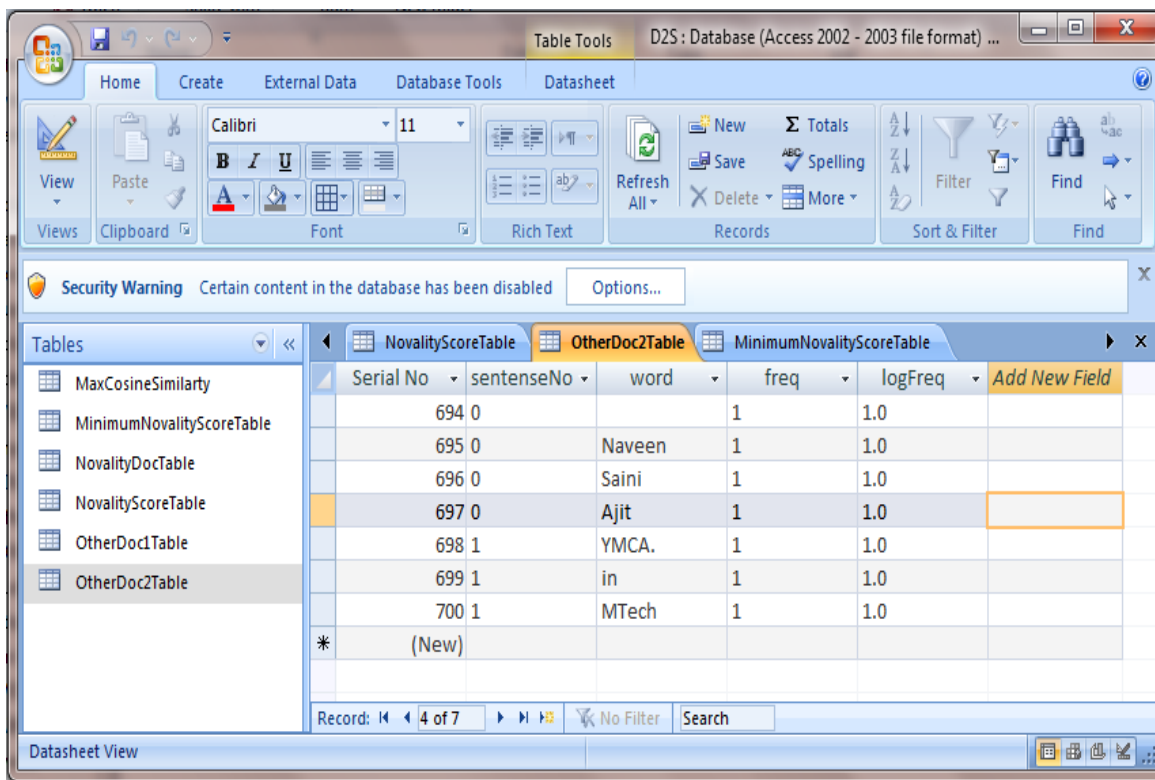


Figure 3.22: Document 2 (History Document)

Security Warning: Certain content in the database has been disabled. Options...

Serial No	sentenceNo	word	freq	logFreq	Add New Field
570	0	Jangra	1	1.0	
571	0	Ajit	2	1.30102999566:	
572	1	bhiwani	1	1.0	
573	1	Bhiwani	4	1.60205999132:	
(New)					

Record: 1 of 4 | No Filter | Search

Figure 3.23: Generated Assumed Novelty Document

Security Warning: Certain content in the database has been disabled. Options...

Serial No	novalityDocSen	otherDoc1	otherDoc2	Add New Field
36	0	-0.60940737306	-0.60940737306	
37	1	1.0	1.0	
(New)				

Record: 1 of 2 | No Filter | Search

Figure 3.24: Novelty Score Computation

Security Warning: Certain content in the database has been disabled. Options...

Serial No	novalityDocSen	otherDoc1	otherDoc2
45	0	1.60940737306	1.60940737306
46	1	0.0	0.0
(New)			

Record: 1 of 2

Figure 3.25: Maximum Cosine Similarity Value

Security Warning: Certain content in the database has been disabled. Options...

Serial No	novalityDocSen	otherDoc1	otherDoc2
24	0	-0.60940737306	-0.60940737306
(New)			

Record: 1 of 1

Figure 3.26: Minimum Novelty Score Value

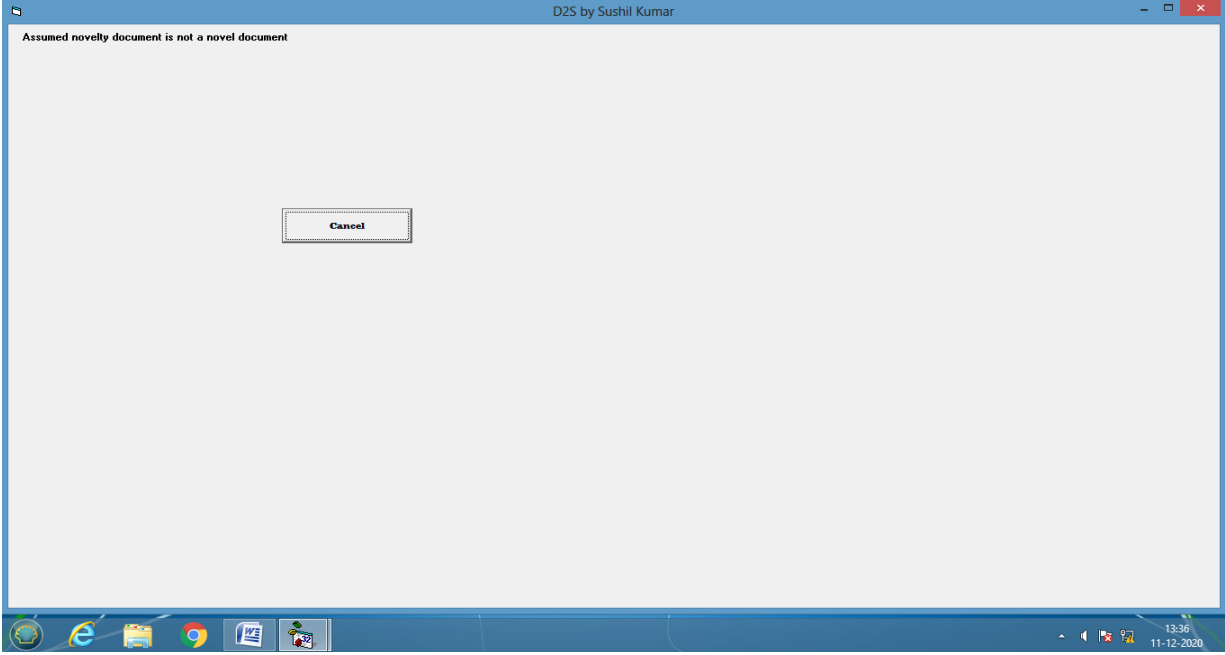


Figure 3.27: Novelty Detection Output

3.5.1 Performance Evaluation

The proposed approach is compared with General Document Retrieval in terms of Precision, Recall and F-Score are defined as:

$$Precision = \frac{\text{Relevant documents Detected}}{\text{Total number of retrieved documents}} \quad (3.4)$$

$$Recall(R) = \frac{\text{Novel documents detected}}{\text{Total documents in data set}} \quad (3.5)$$

$$F - Score = \frac{2 * P * R}{P + R} \quad (3.6)$$

Table 3.13 lists out the precision, recall and F-score values for keyword similarity based on general document retrieval and proposed technique:

Table 3.14: Precision, Recall and F-Score Comparison

Query	Count of Documents Considered	General Document Retrieval			Proposed D2S Approach		
		Precision	Recall	F-Score	Precision	Recall	F-Score
Diwali	30	0.16	0.5	0.24	0.23	0.7	0.34
Holi	30	0.1	0.42	0.16	0.13	0.57	0.20
Cricket	30	0.13	0.5	0.2	0.166	0.62	0.26
Pulwam a Attack	30	0.2	0.66	0.3	0.23	0.77	0.35

The precision, Recall and F-Score plots are shown in figures 3.28, 3.29 and 3.30 respectively

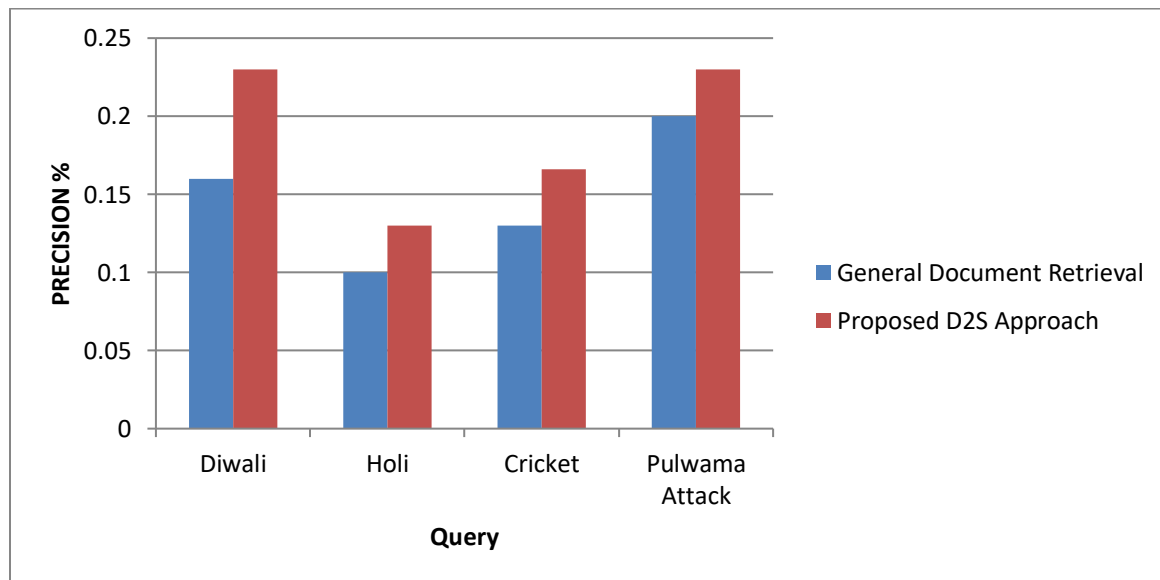


Figure 3.28: Precision Plot

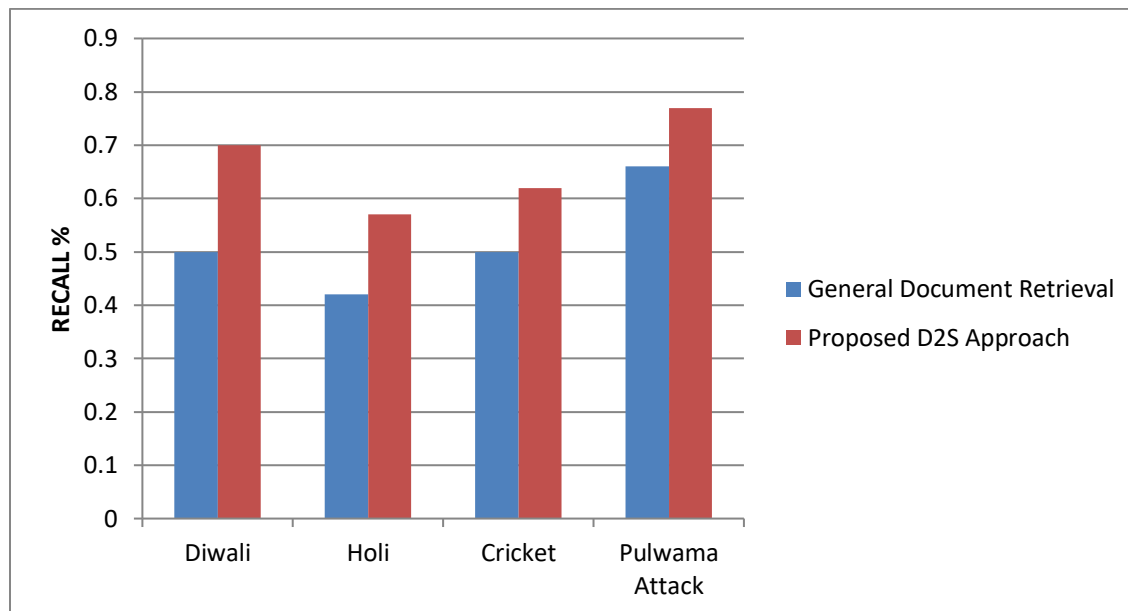


Figure 3.29: Recall plot

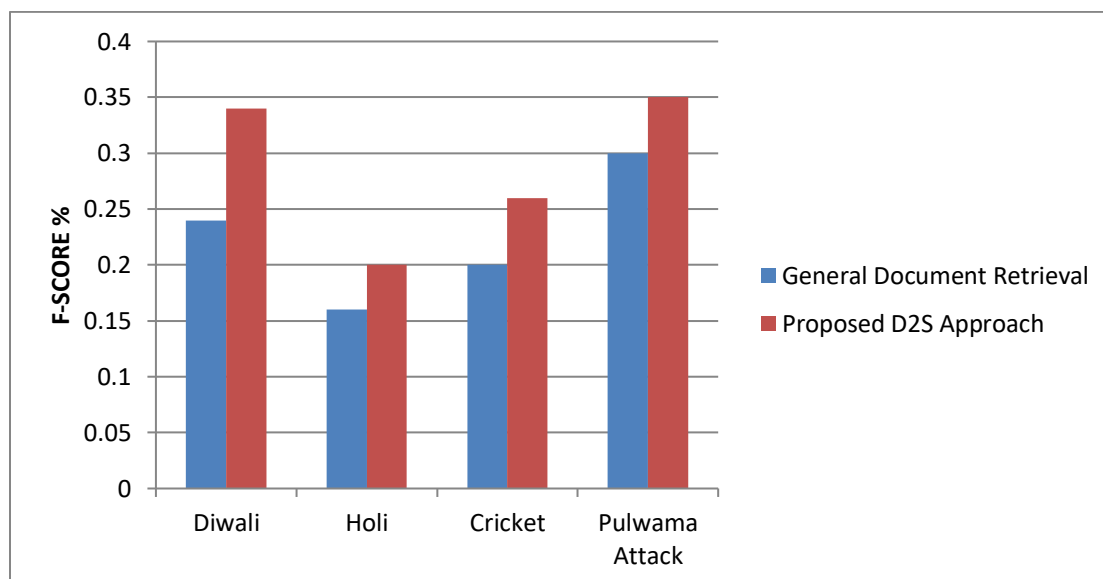


Figure 3.30: F-Score plot

From the implementation it has been proved that this proposed method gives proper result in lesser amount of time and with better efficiency. The precision, recall and F-measure are also high when this algorithm applied on the set of documents as compared with general document retrieval system

Table 3.15 Comparison with General Novelty Detection work

General Novelty Detection	Proposed Approach
<ul style="list-style-type: none">• Earlier approaches treated the sentences and documents as two distinct resources and find novelty separately.	<ul style="list-style-type: none">• The proposed approach tells a document as redundant if it shares a single sentence with the previously seen document.• The work mainly enhances the code reuse for novelty identification as it concerned on sentence level.

Above table 3.14 shows the comparison between the proposed approach for novelty identification and the approaches in the literature.

3.6 SUMMARY

The work proposed a framework which can be applied on document level novelty detection. This framework can make document-level novelty detection more effective by adopting the techniques for the sentence level. Experiments with different examples show that proposed method can greatly improve the document level novelty detection performance in terms of precision and recall and F-score. Therefore, the proposed work may be used to incorporate in to real world information retrieval system.

CHAPTER IV

TEXT SUMMARIZATION BASED NOVELTY DETECTION

4.1 INTRODUCTION

Expanding development of data volume on the web makes an expanding need grow new automatic techniques for recovery of text documents having novel sentences and positioning them as per their pertinence to the client query. Essentially the web index recovers the greater part of data about text documents having duplicate sentences too. This makes issue of finding the significant data and furthermore the refreshed data. So it isn't easy to use. To make it more valuable and to determine the issue of duplicate sentences and information the thought "Novelty detection utilizing text Summarization" is by all accounts more suitable and gainful.

The work is to limit the repetitive sentences from a text document in such a way that it gives the significant and novel sentences as it were. Different methodologies, for example, the single use of text summarization [60, 61] or general novelty detection may not give the novel sentence or data. So this work is a consolidated methodology of both text summarization and general novelty identification procedure.

4.2 TEXT SUMMARIZATION

It is a method to reduce the content of text document, which makes a summary that holds the significant data in the original document. Text summarization [105], eliminate the unessential information as well as the duplicate data too. As the issue of information over-burden has become over the web, so the interest in the content summarization helps in lessening the volume of information. This method included a single text summarization or multi document outline. Summarization has been decayed into three principle stages:

- Source text translation in to a text representation
- Text representation is changed in to summary representation.
- From summary representation to generate the summarize document

The method of text summarization [107] diminishes a text in to a section or passage that passes on the significant purposes of the content. The looking of significant data from a large stream of documents is exceptionally troublesome job for the users. Consequently the automatic text extraction of data or summary of the content document is required. This text summarization causes the user to decrease time to peruse the entire document and it gives needed information from the large volume.

With the fast development of the World Wide Web (WWW), data over-burden is turning into an issue for getting to the necessary data. Text outline can be an irreplaceable solution for decrease the volume and give the required information

4.2.1 Types of Text Summarization

The process of text summarization classify in to two types:

- Abstractive
- Extractive

Abstractive text summarization: It constructs an interior semantic representation and afterward utilizes common language generation strategy to make a summary that is nearer to what a human may create. It is retelling something very similar in different words.

Extractive Text Summarization: This selects the important sentences from the text by using word frequency, cue word etc. and generate the summary.

In this work the extractive text summarization is used to generate the summary of the text document.

4.2.2 Extractive Text Summarization Performs Two Levels:

- **Pre Processing:** It is an essential step to load content in to the proposed framework and different tasks are performed on the document, for example, boundary reorganization, stop word removal and stemming and so on.
- **Processing:** In this process the features impact the significance of sentences are chosen and determined. At that point the weights are allocated utilizing learning strategies and high level sentences are chosen.

4.3 TERM FREQUENCY

In information recovery, TF-IDF, short for term frequency-inverse document frequency, is a mathematical measurement that is proposed to reflect how significant a word is to a document in a corpus. It is regularly utilized as a weighting factor in information recovery and text mining. The TF-IDF esteem expands relatively to the occasions a word shows up in the document, yet is balanced by the frequency of the word in the corpus, which assists with adapting to the way that a few words show up more frequently in general.

The term frequency is significant component, which speaks to how often the term appear in the document (as a rule a compression function is used, for example, square root or logarithm is applied) to compute the term frequency. The term identifying the sentence boundaries in a document is based on punctuation and split into sentences.

4.3.1 Keyword Frequency

The keywords are the top high frequency words in term sentence frequency. After cleaning of document the frequency of each word is calculated. The word score are chosen as keywords and based on this feature, any sentence in the document is scored by number of keywords it contains.

4.4 INVERSE DOCUMENT FREQUENCY (IDF)

The inverse document frequency is a measure of how much information the word provides that is, whether the term is common or rare across the document. It is the logarithmically scale inverse function of the document of the document that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$TFIDF = (1 + \log_{10} N) * IDF$	(4.1)
-----------------------------------	--------------

Where, N = number of terms, the value for IDF =1.

If , the value of N=0 than the TFIDF= 0

4.4.1 Algorithm for TF-IDF

<p>Step 1: Take source document as input</p> <p>Step 2: Done preprocessing on source document</p> <p>Step 3: Extract the term frequencies</p> <p>Step 4: for all the terms</p> <p> 4.1 generate the TFIDF value</p> <p>Step 5: for all TFIDF value</p> <p> 5.1 Extract the sentences from source text document the terms with highest TFIDF values</p> <p>Step 6: provide a summary for source document with sentences extracted by step 5.1</p> <p>Step 7: final output is summarized document.</p>
--

Figure 4.1: Algorithms for TF-IDF

4.5 PROPOSED ARCHITECTURE FOR TEXT SUMMARIZATION BASED NOVELTY DETECTION

The Proposed architecture as shown in figure 4.2 takes the source text as input. Then text summarization method is applied on the input text document. On the summarized text document the novelty detection technique is applied. The output of this document is the extracted novel sentences. This document is compared with the document on which the general novelty method was applied. This method involves the following steps:

- (a) Text Summarization method applied on the text document which provides the summary.
- (b) Novelty Detection method is applied on the summarized text document.
- (c) It provides the novel sentences by using cosine similarity method.
- (d) The final output is compared with the original document

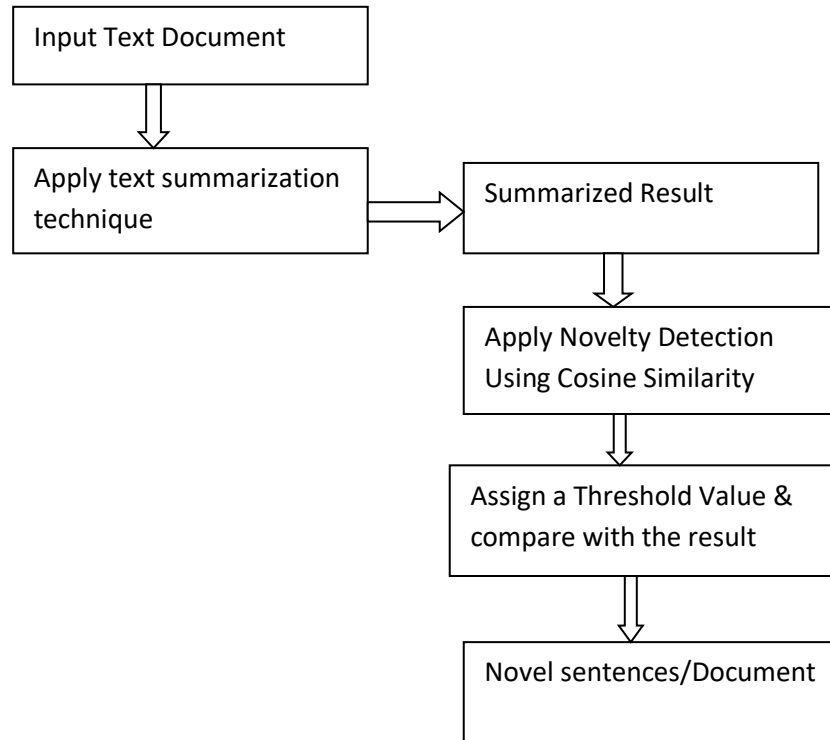


Figure 4.2: Proposed Architecture

Then text summarization method is applied on the input text document. On the summarized text document the novelty detection technique is applied. The output of this document is the extracted novel sentences.

4.5.1 ALGORITHM FOR PROPOSED TECHNIQUE

The figure 4.3 shows the algorithm for text summarization based novelty detection which includes the sequence of steps applied to determine the novelty of documents.

Step 1: Take Source document as input D_i

Step 2: Done pre-processing on D_i

Step 3: Extract term frequency (TF) for all terms(T) of D_i

Step 4: Estimate TFIDF for all T

Step 5 : Choose all T with maximum TFIDF value and extract all sentences with such T

Step 6: Generate the Summary (Si) of Di

Step 7: Splits Si (where i=1 to n) into sentences (S1, S2,S3,.....Sn) each sentences as an individual document SDi (where i=1 to n)

Step 8: Find all TF for each T in the SDi

Step 9: Calculate the cosine similarity value for each SDi

$$\text{Similarity} = \cos(\theta) = \frac{SDi \cdot SDi+1}{||SDi|| ||SDi+1||} = \frac{\sum_{i=1}^n SDi \times SDi+1}{\sqrt{\sum_{i=1}^n (SDi)^2} \times \sqrt{\sum_{i=1}^n (SDi+1)^2}}$$

Step 10: Set the threshold value for all SDi

If similarity > threshold

Then SDi is Novel (Ni)

Else not Novel

Step 11: If Ni has some mutual or intersected COS Θ value than minimize Ni(for i=1 to n)

Step 12: Final NDi (i=1 to n) novel sentences for Di are generated.

Figure 4.3: Algorithm for Proposed Technique

4.6 Implementation and Results

The approach exploited on a set of documents. This approach takes the text as input and after stop word removal, lemmatization it gives the extractive summary as output text. Then cosine novelty method applied on the summarized text document. The output document then compared with the document on which normal text summarization was applied.

4.6.1 Experimental Result 1

The original text document used for basic analysis is shown in figure. 4.4 and the Cosine similarity function is applied on this.

Ymca is a university. Ymca is a government university. Ymca is a government university of Haryana. Ymca university is a top university of Haryana. I live in ymca hostel. Ymca is a university of science and technology.

Figure 4.4: Original Document

4.6.1.1 Cosine Similarity Calculation

This document is segmented in to the sentences and suppose these sentences A,B,C,D,E and F acts as individual documents.

Step 1: Split the document in to sentences as shown below:

- A) Ymca is a university.
- B) Ymca is a government university.
- C) Ymca is a government university of Haryana.
- D) Ymca university is a top university of Haryana.
- E) I live in ymca hostel.
- F) Ymca is a university of science and technology.

Step 2: Find the term frequency for each document as shown in Table 4.1.

Table 4.1: Term Frequency Values of Original Document

TERMS	TERMS FREQUENCY FOR DOCUMENTS					
	A	B	C	D	E	F
Ymca	1	1	1	1	1	1
university	1	1	1	2	0	1
governement	0	1	1	0	0	0
haryana	0	0	1	1	0	0
Top	0	0	0	1	0	0
Live	0	0	0	1	1	0
Hostel	0	0	0	1	1	0
science	0	0	0	1	0	1
technology	0	0	0	0	0	1

Step 3: After that the Cosine similarity function is used to find the similarity between the two documents. The eqn. 4.2 shows the cosine similarity function:

$$\text{Similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (4.2)$$

In the table 4.1, six documents are involved. All the documents are compared with each other and the total number of comparison is calculated by

$$\text{No. of Comparisons} = \frac{n(n-1)}{2} \quad (4.3)$$

Where, $n = 6$ and the total number of comparison are 15.

Step 4: Calculation of Cosine values for each comparison

For example for the sentence (a):-

$$\text{Cos}(a,b) = 1 \cdot 1 + 1 \cdot 1 / \sqrt{((1^2 + 1^2) * (1^2 + 1^2 + 1^2))} = 0.81$$

$$\text{Cos}(a,c) = 1 \cdot 1 + 1 \cdot 1 / \sqrt{((1^2 + 1^2) * (1^2 + 1^2 + 1^2 + 1^2))} = 0.70$$

$$\text{Cos}(b,c) = 1 \cdot 1 + 1 \cdot 1.30 / \sqrt{((1.30^2 + 1^2 + 1^2 + 1^2) * (1^2 + 1^2))} = 0.87$$

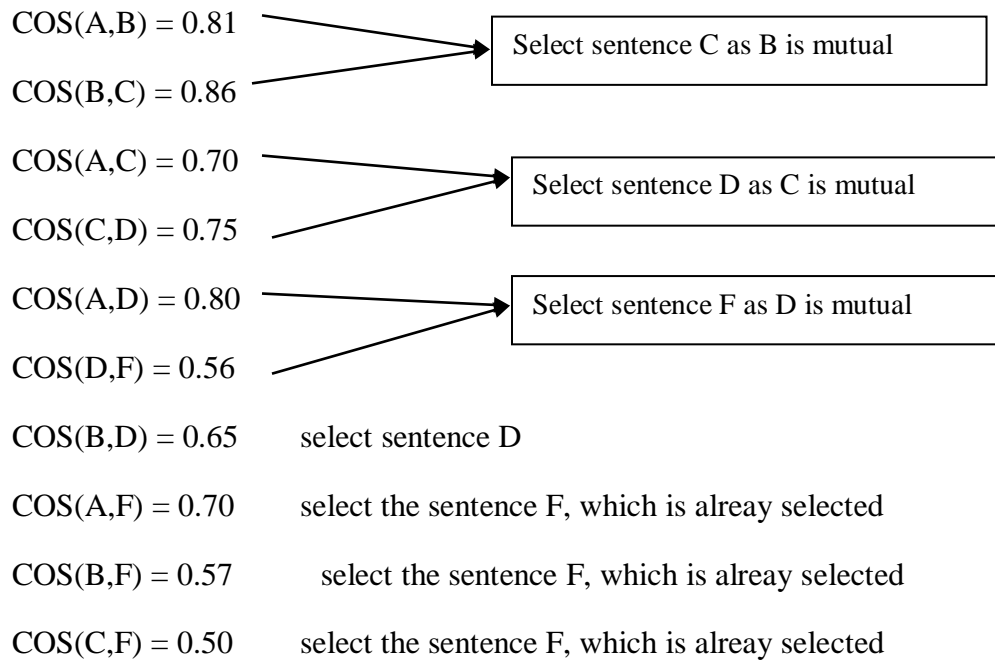
Likewise find all values of $\cos \theta$ for other pairs. Table 4.2 reflects Cosine values of all the possible pairs.

Table 4.2: Cosine Values

Comparison	COS θ Value	Comparison	COS θ Value	Comparison	COS θ Value
(A,B)	0.81	(B,C)	0.86	(C,E)	0.28

(A,C)	0.70	(B,D)	0.65	(C,F)	0.50
(A,D)	0.80	(B,E)	0.33	(D,E)	0.21
(A,E)	0.40	(B,F)	0.57	(D,F)	0.81
(A,F)	0.70	(C,D)	0.75	(E,F)	0.28

From the table 4.2 value in bold taken out, which are greater than the threshold. The threshold value used is 0.45 and the pairs which are having such values are:



So that only 3 sentences C,D,F are selected and included into the summarized document.

Thus the novel sentence extracted are:

Ymca is a government university of Haryana. Ymca university is a top university of Haryana. Ymca is a university of science and technology.

Figure 4.5: Extracted Novel Sentences

4.6.1.2 Apply the TF-IDF Technique on the Original Document:

Ymca is a university. Ymca is a government university. Ymca is a government university of Haryana. Ymca university is a top university of Haryana. I live in ymca hostel. Ymca is a university of science and technology.

Figure 4.6: Document to apply TF-IDF Method

Step 1: Find the TFIDF of the original document by using $TFIDF = (1 + \log_{10} N) * IDF$ and choose IDF=1.

Table 4.3: TF-IDF Value

Term	TF	TFIDF
YMCA	6	1.77
University	6	1.77
Government	2	1.30
Haryana	2	1.30
Top	1	1
Live	1	1
Hostel	1	1
Science	1	1
Technology	1	1

Step 2: The bold values show the highest TF-IDF value, so extract all sentences having terms YMCA and University

Step 3: Summarized document after TF-IDF

Ymca is a university. Ymca is a government university. Ymca is a government university of Haryana. Ymca university is a top university of Haryana. I live in ymca hostel. Ymca is a university of science and technology.

Figure 4.7: Summarized document

From the fig. 4.7 TF-IDF provided the same original document as a summarized document. So it is the disadvantage of TF-IDF.

4.6.1.3 Apply the Cosine Similarity on the Summarized Data

Step 1: Calculate TF for each document A,B,C,D,E,F from the fig. 4..6. The terms frequency reflected into the table 4.4

Table 4.4: Term Frequency values after summarization

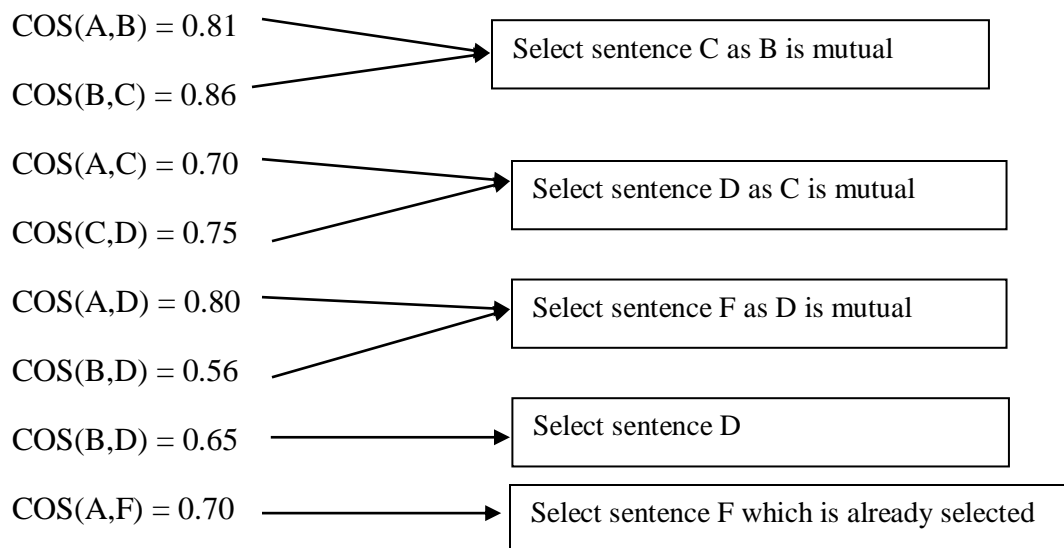
TERMS	TERMS FREQUENCY FOR DOCUMENTS					
	A	B	C	D	E	F
ymca	1	1	1	1	1	1
university	1	1	1	2	0	1
gouvernement	0	1	1	0	0	0
haryana	0	0	1	1	0	0
top	0	0	0	1	0	0
live	0	0	0	1	1	0
hostel	0	0	0	1	1	0
science	0	0	0	1	0	1
technology	0	0	0	0	0	1

Step 2: The total comparisons involved to compare the Cosine values are $6(6-1)/2=15$.

Table 4.5: Cosine Values

Comparision	COS Θ Value	Comparision	COS Θ Value	Comparision	COS Θ Value
(A,B)	0.81	(B,C)	0.86	(C,E)	0.28
(A,C)	0.70	(B,D)	0.65	(C,F)	0.50
(A,D)	0.80	(B,E)	0.33	(D,E)	0.21
(A,E)	0.40	(B,F)	0.57	(D,F)	0.81
(A,F)	0.70	(C,D)	0.75	(E,F)	0.28

From the table 4.5 shaded the value which are greater than the threshold. The threshold value used is 0.45. the pairs which are having such values are:



Thus the novel sentence extracted are:

Ymca is a government university. Ymca university is a top university of Haryana. Ymca is a university of science and technology.

Now observe that TF-IDF value for term YMCA and UNIVERSITY is higher. So choose the sentences having YMCA and UNIVERSITY terms. It gives the summarized document of the original document.

Step 3: Figure 4.7 shows the summarization of the original document

Ymca is a university. Ymca is a government university. Ymca is a government university of Haryana. Ymca university is a top university of Haryana. I live in ymca hostel. Ymca is a university of science and technology.

Figure 4.8: Summarized document

TFIDF provided the same text document as is original text.

This is a disadvantage that it may retrieve the same original document as a summarized document. Hence to resolve this disadvantage the use cosine similarity together with summarization method to extract the novel sentences..

4.6.1.4 TF-IDF Plus Cosine Similarity Approach

Step 1:- Break the text document retrieved after TF-IDF procedure into segmentation of sentences like:-

- Ymca is a university.
- Ymca is a government university.
- Ymca is a government university of Haryana.
- Ymca university is a top university of Haryana.
- I live in ymca hostel.
- Ymca is a university of science and technology.

Step 2:- Calculate term frequency of each individual sentence as shown in below table 4.6.

Find TFIDF value for each using formula:- $TFIDF = (1 + \log_{10} N) * IDF$

Take IDF value= 1 always

Do not compute value when term frequency = 0 take it as it is i.e = 0

Table 4.6 TF-IDF Values after TF-IDF and Cosine Approach

Term	Frequency						TF-IDF					
	a	b	c	d	e	f	a	b	c	d	e	f
ymca	1	1	1	1	1	1	1	1	1	1	1	1
university	1	1	1	2	0	1	1	1	1	1.30	0	1
government	0	1	1	0	0	0	0	1	1	0	0	0
haryana	0	0	1	1	0	0	0	0	1	1	0	0
top	0	0	0	1	0	0	0	0	0	1	0	0
live	0	0	0	0	1	0	0	0	0	0	1	0
hostel	0	0	0	0	1	0	0	0	0	0	1	0
science	0	0	0	0	0	1	0	0	0	0	0	1
technology	0	0	0	0	0	1	0	0	0	0	0	1

Now apply the cosine formula over TF-IDF values got in table 4.6.

$$\text{Similarity} = \cos(\theta) = \frac{A.B}{||A|| ||B||} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (4.3)$$

For example for the sentence (a):-

$$\text{Cos(a,b)} = 1*1 + 1*1 / \sqrt{((1^2 + 1^2) * (1^2 + 1^2 + 1^2))} = 0.81$$

$$\text{Cos(a,c)} = 1*1 + 1*1 / \sqrt{((1^2 + 1^2) * (1^2 + 1^2 + 1^2 + 1^2))} = 0.70$$

$$\text{Cos(b,c)} = 1*1 + 1*1.30 / \sqrt{((1.30^2 + 1^2 + 1^2 + 1^2) * (1^2 + 1^2))} = 0.87$$

Likewise we find all values of $\cos \theta$

Table 4.7 summarizes all the values of the possible pair for cosine function.

Table 4.7 Threshold Comparison

Cosine similarity values of each sentences	Cosθ value	After threshold value application
Cos(a,b)	0.81	0.81>0.45
Cos(a,c)	0.70	0.70>0.45
Cos(a,d)	0.87	0.87>0.45
Cos(a,e)	0.40	0.40<0.45
Cos(a,f)	0.70	0.70>0.45
Cos(b,c)	0.86	0.86>0.45
Cos(b,d)	0.71	0.71>0.45
Cos(b,e)	0.33	0.33<0.45
Cos(b,f)	0.57	0.57>0.45
Cos(c,d)	0.62	0.62>0.45
Cos(c,e)	0.28	0.28<0.45
Cos(c,f)	0.50	0.50>0.45

Cos(d,e)	0.26	0.26<0.45
Cos(d,f)	0.62	0.62>0.45
Cos(e,f)	0.28	0.28<0.45

Now provide a threshold value= .45 and the sentences are selected with $\text{COS}\Theta > \text{threshold value}$. Hence all the sentences with cosine value less than threshold are eliminated.

From cos θ values selected values are:

Cos(a,b):- 0.81 sentence (b) is selected

Cos(a,c):- 0.70 sentence (c) is selected

Cos(a,d):- 0.87 sentence (d) is selected

Cos(a,f):- 0.70 sentence (f) is selected

Cos(b,c):- 0.86 sentence (c) is selected

But sentence (b) is consisted in (c) from last value so drop sentence (b).

So, the final novel sentences in the document are:-

Ymca is a government university of Haryana. Ymca university is a top university of Haryana. Ymca is a university of science and technology.

Figure 4.9: Extracted Novel Sentences

So, the proposed technique gives better result than the existed one.

4.6.2 Example 2:

Text document in figure 4.8 is used for basic analysis:

Ram is writing .He is writing with a pen. The pen is a blue pen. He is writing a letter with a blue pen.

Figure 4.10: Document 2

Step 1: Calculate term frequency after stemming, lemmatization and stop word removal.

Ram write write pen pen blue pen write letter blue pen

Figure 4.11: Document 2 after Preprocessing

Find TFIDF value for each using formula:- $TFIDF = (1 + \log_{10} N) * IDF$

Take IDF value= 1 always

Do not compute value when term frequency = 0 take it as it is i.e. = 0

Step: 2: Now construct a TF and TF-IDF values:

Table 4.8 TF and TF-IDF Values Example 2

Terms	TF	TF-IDF
Ram	1	1
Write	3	1.47
Pen	4	1.60
Blue	2	1.30
Letter	1	1

The value 1.60 in table 4.8 is showing the highest TF-IDF value means the occurrence of term PEN is high. So, extract a summary of sentences having the term Pen and drop all sentences not having term Pen.

Thus the summary of original document after TF-IDF operation is shown in figure 4.10

He is writing with a pen. The pen is a blue pen. He is writing a letter with a blue pen.

Figure 4.12: Summarized Document 2

Hence, the summary consists of only 3 sentences by eliminating 1 sentence.

- **Cosine Similarity Calculation over the result provided by the TF-IDF:**

Step 1:- Segmentation of sentences is done of the summarized document.

- | |
|---|
| a) He is writing with a pen.
b) The pen is a blue pen.
c) He is writing a letter with a blue pen. |
|---|

Figure 4.13: Document 2 for Proposed Technique

Find TF-IDF value for each using formula:- $1 + \log \text{frequency} * \text{IDF}$

Take IDF value= 1 always

Do not compute value when term frequency = 0 take it as it is i.e. = 0

- **Now calculate the TF-IDF for all sentences**

Table 4.9: TF and TF-IDF Values after Summarization

Term	Frequency			TF-IDF		
	a	b	c	a	b	c
Write	1	0	1	1	0	1
Pen	1	2	1	1	1.30	1
Blue	0	1	1	0	1	1
Letter	0	0	1	0	0	1

Now apply the cosine formula over TF-IDF values using eqn. 4.4

$\text{Similarity} = \cos(\theta) = \frac{A.B}{ A B } = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad (4.4)$

For example for the sentence (a):-

$$\text{Cos}(a,b) = 1*0 + 1*1.30 / \sqrt{((1^2 + 1^2) * (1.30^2 + 1^2))} = .56$$

$$\text{Cos}(a,c) = 1*1 + 1*1 / \sqrt{((1^2 + 1^2) * (1^2 + 1^2 + 1^2))} = .707$$

$$\text{Cos}(b,c) = 1*1 + 1*1.30 / \sqrt{((1.30^2 + 1^2) * (1^2 + 1^2 + 1^2))} = .701$$

Now from these $\text{cos}(a,b)$ and $\text{cos}(a,c)$ are selected from these values sentence (b) and (c) are selected. But from $\text{cos}(b,c)$ sentence (b) gets eliminated.

Hence, the summarized text document consist only sentence (c). Hence, this method provides more efficient result as compared to the TF-IDF technique.

Final summarized text document is:

He is writing a letter with a blue pen.

Figure 4.14: Extracted Novel Sentence

From the example 2, it has need clear that the proposed technique provides novel sentence as compared with general novelty detection using cosine similarity.

4.7 SUMMARY

From the above examples, it has been cleared that the technique TF-IDF for text summarization may retrieve the same original text document as a summarized document. TFIDF provides poor result for text summarization. To overcome this problem, the proposed method has been used the cosine similarity function in combination with the TF-IDF technique. Cosine similarity involves more number of the document comparison than the proposed Technique and having a more time and space complexity. TF-IDF first minimize original document as summarize document and reducing the number of sentences. Thus the proposed approach has less number of documents comparisons than the cosine approach.

Hence, it has been concluded from the examples that the proposed technique provided the better result to extract novel sentences.

CHAPTER V

CLUSTERING BASED NOVELTY DETECTION IN TEXT DOCUMENTS

5.1 INTRODUCTION

Novelty detection is a technique used to retrieve relevant and novel information according to the user query with less effort. The goal is for the user to quickly get useful information without going through a lot of redundant information, which involves more effort and time. This chapter proposed a clustering [82] based approach for novelty detection which provides the relevant and novel information for the user query. Clustering is a method used to collect the similar documents in a cluster. In this approach the incoming stream of documents has been clustered using k-means algorithm [83] according to the given query. Then the cluster heads are selected from the various clusters having minimum distance within a cluster. These cluster heads are the novel documents from a collection of documents according to the given query.

5.2. PROPOSED METHODOLOGY FOR CLUSTERING BASED NOVELTY DETECTION

The proposed work used a clustering based approach for novelty detection which provides the relevant and novel information to the user query. Firstly the incoming stream of documents based on user query related to a given domain clustered using k-means algorithm [84,85,86]. User makes a query based on specific domain using search engine and the first thirty retrieved results scraped out and store in a file on the disk. These documents are used to make ten clusters each containing of similar documents. Based on these clusters one cluster head is selected from each cluster which provides ten documents. All of these ten documents having large distance as compared to with each other. This provided a list of ten novel documents based on the query. The system consists of various modules i.e user, search engine, results retrieved; scarp results file storage and finally the novelty detection module. The user can provide the query to the search engine interface and the search engine provide the list of retrieved results. Afterward these retrieved results are scraped out in a document format and stored on disk as CSV (Comma Separated values) file. On this file the novelty detection algorithm based on clustering has been applied. The architecture comprises with these modules shown in figure 5.1.

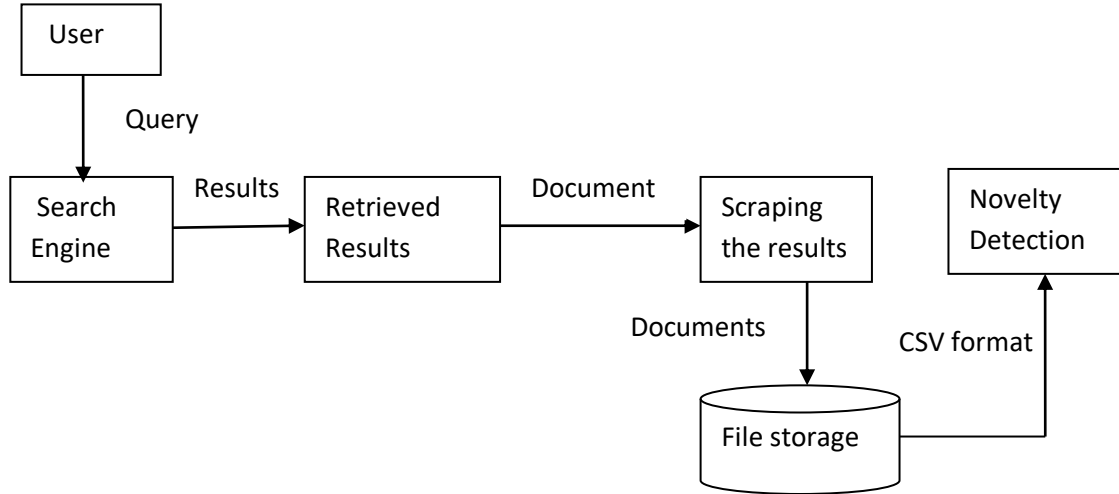


Figure 5.1: Architecture for Clustering Based Novelty Detection

Initially the experiment has been performed on a set of thirty documents related to different domains based on the query. Thirty documents read from CSV file that was a clean data. Then TF-IDF *vectorizer* is used to convert in to TF-IDF vectors and trained this by using K-means model. The clusters are formed based on the similarity between the documents. The following steps are involved in the process of novel documents extraction as follow:

5.2.1 Collection of Textual Datasets

At the beginning one hundred and fifty documents were collected which consists of thirty each of festival (f1,f2.....f30),sports (s1,s2.....s30), technology (t1,t2....t30), politics (p1,p2.....p30) and education domains (e1,e2...e30). These documents undergo refinement which is fed to the algorithm to obtain clusters containing documents from similar domains.

5.2.2 Convert documents into Vectors

A document comprises of countless sentences, so the archive is parts into sentences. The sentences comprise of an enormous number of words, and it isn't generally fundamental that each word is of significance. Because of which high dimensionality of the record must be decreased by handling the report to dispose of additional words to get the heaviness of every one of the word to be utilized in the calculation. The conversion of documents into vectors is carried in various steps.

5.2.2.1 Tokenization

The processes involve in information retrieval require the words of documents. Tokenization is used to identify the meaningful words called tokens. The main use of tokenization is to split the sentences into individual tokens. For example 'Ram is eating apple' so in this sentence 'Ram', is, 'eating', and apple' are the tokens.

5.2.2.2 Stop-words Removal

Stop words are much of the time happening, irrelevant words that show up in a data set record, article, or a page, and so forth pronoun, qualifier, relational word and so on which are utilized all through in the record must be taken out to get appropriate outcome. For example 'Can Tom laughing' so after removal of stop word can and be it results in Tom, laughing.

5.3. NOVELTYDETECTION MODULE

This module helps in finding the novelty of the documents. The document uses to apply the algorithm is shown in figure 5.2

Education is a pillar of development. Educated citizens help in developing a nation. India is a developing nation. Indian education system will help the nation to develop. Various schemes have been launched to motivate the citizens to educate. Due to these schemes literacy rate is increasing. One day when India will reach to its goal of education it will definitely become developed nation.

Figure 5.2: Document for basic analysis

Firstly text documents splits in to sentences and each sentence is act as a document like as $d1, d2, \dots, dn$. Then pre-processing the document using tokenization, remove stop words, replace tokens by their stems and generate inverse document frequencies vectors on dynamic vector space model. Then the system picked up relevant documents for a given query and filter out the non-relevant documents in the categorization stage. Finally based on the historical documents, the system determined whether the input document is novel or not.

Figure 5.3 show the algorithm for K-means clustering.

The algorithm takes the following steps:

Input: X, K where X=Set of classified instances, K= integer, Output: Set of K clusters

Require $X \neq \text{Null}$, $K > 0$

1. Procedure GenerateClusters
2. Initialize K random centroids
3. Repeat
4. for all instance i in X do
5. shortest $\leftarrow 0$
6. membership $\leftarrow \text{null}$
7. for all centroid c1 do
8. dist1 $\leftarrow \text{distance}(c1)$
9. if dist1 < shortest then
10. shortest $\leftarrow \text{dist1}$
11. membership $\leftarrow c1$
12. end if
13. end for
14. end for
15. Recalculate Centroid(c1)
16. Until convergence
17. End procedure

Figure 5.3: Algorithm for K-means clustering

Let us assume $K=2$ and the D5 and D7 are chosen for clusters

Calculate Euclidean distance using the given equation

$$\text{Distance}[(x, y), (a, b)] = \sqrt{(x - a)^2 + (y - b)^2} \quad (5.1)$$

The distance matrix with term frequency in documents is shown in table 5.1

Table 5.1: Distance matrix with Term Frequency

Terms/documents	D1	D2	D3	D4	D5	D6	D7
Education	1	1	0	1	1	0	1
Pillar	1	0	0	0	0	0	0
Development	1	1	1	0	0	0	1
Citizen	0	1	0	0	1	0	0
Nation	0	1	1	1	0	0	1
India	0	0	1	1	0	0	1
System	0	0	0	1	0	0	0
Various	0	0	0	0	1	0	0
Schemes	0	0	0	0	1	0	0
Launch	0	0	0	0	1	0	0
Motivate	0	0	0	0	1	0	0
Literacy	0	0	0	0	0	1	0
Increase	0	0	0	0	0	1	0
Reach	0	0	0	0	0	0	1
Goal	0	0	0	0	0	0	1
Definitely	0	0	0	0	0	0	1

Euclidean Distance between D1 and D5 is calculated by using equation no. 5.1

$$\sqrt{(1-1)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2}$$

$$\sqrt{0+1+1+1+0+0+0+1+1+1+1+0}$$

$$= \sqrt{7}$$

$$= 2.67$$

Similarly, the distance can be calculated with the other documents.

The table 5.2 shows the movement of documents D1, D2, D3, D4, D7 move to cluster D7 and D5, D6 move to cluster D5 based upon the minimum Euclidean distance:

Table 5.2: Movement of Clusters

Document	D5 Cluster	D7 Cluster	Minimum distance	Movement to cluster
D1	2.64	2.44	2.44	D7
D2	2.64	2.23	2.23	D7
D3	3	2	2	D7
D4	2.82	2	2	D7
D5	0	3.31	0	D5
D6	2.82	3	2.82	D5
D7	3.31	0	0	D7

The CSV clean data file trained by using K-means [80, 81] model and clusters of documents are formed. Then find out the cluster head from K-means model in document format from each cluster. These clusters heads from each cluster are having large distance with other cluster head. Therefore, the collection of these cluster heads yields the novel documents from a collection of thirty documents.

5.4 IMPLEMENTATION OF CLUSERING BASED NOVELTY DETECTION

Initially the experiment has been performed on a set of thirty documents related to different domains based on the query. Thirty documents read from CSV file that was a clean data. Then TF-IDF *vectorizer* is used to convert in to TF-IDF vectors and trained this by using K-means model. The clusters are formed based on the similarity between the documents. From these clusters the cluster head are formed from K-means model in the document format. These clusters heads from each cluster are the novel documents from a collection of thirty documents according

to the query. The implementation included Anaconda (Python distribution) *Jupyter* Notebook Python 3.6 [90] is a free open source distribution of the Python and R-programming languages. The use of this as it provide support for scientific computing i.e data science, machine learning applications, large scale data processing and predictive analytics that aims to simplify package management and deployment. It also include the necessary library *Numpy* for numerical calculation, *Pandas* for opening large file in hard disk, *Sklearn* for importing K-means model and *nltk* library to clean data. The steps that are used to execute the algorithms are below:

Step 1: Figure. 5.4 used to make a query for Bing search engine.

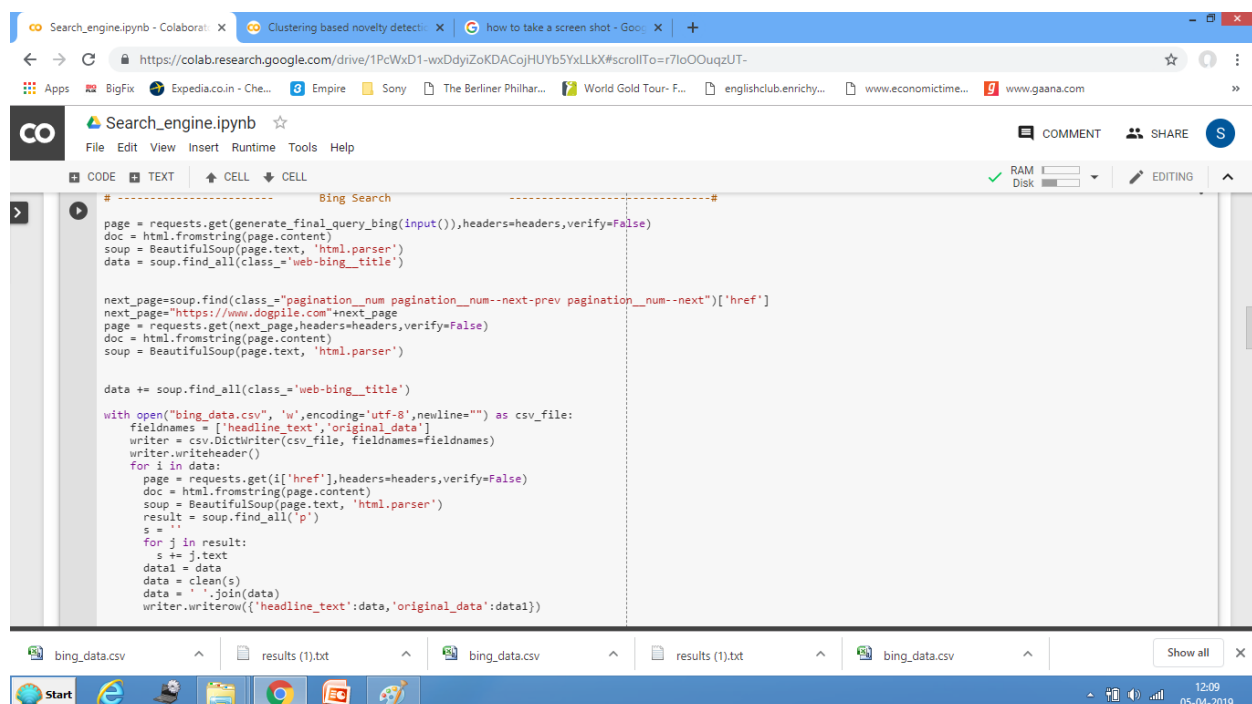
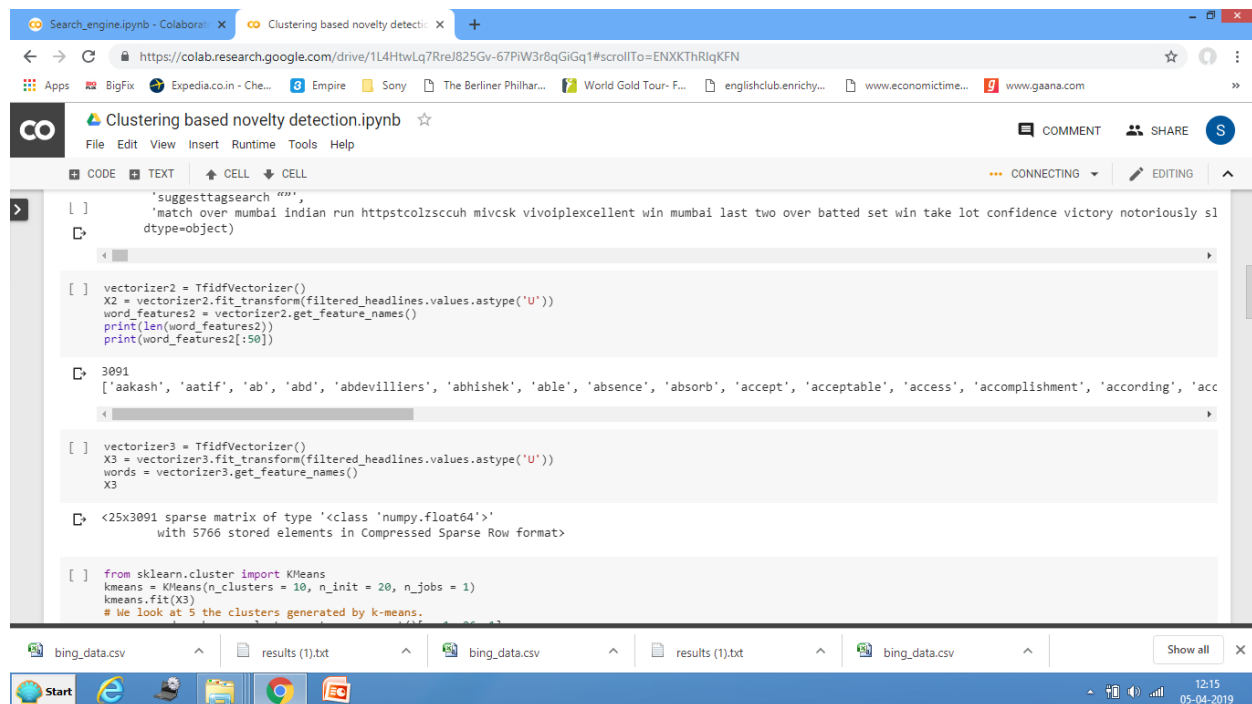


Figure 5.4: Interface for Bing Search Engine

When this module is run by pressing on the arrow sign on top left side of the module, it provides the interface to make a query. User can make a query in any of domains i.e festival, politics, entertainment, sports and education. After making the query for example, ‘holi’ related to domain festival the module first scrap the thirty results from Bing search engine by using *BeautifulSoup* method of *Jupyter* notebook. Then these thirty documents are written in the *bing_data.csv* file with header text data and original data. This file is automatically downloadable on the system.

Step 2: To open the clustering based novelty detection file and perform the necessary refinements with available libraries



```
[ ] 'suggettagsearch ""',
[ ] 'match over mumbai indian run httpcolzscguh mivcsk vivoiplexcellent win mumbai last two over batted set win take lot confidence victory notoriously sl
dtype=object)

[ ] vectorizer2 = TfidfVectorizer()
X2 = vectorizer2.fit_transform(filtered_headlines.values.astype('U'))
word_features2 = vectorizer2.get_feature_names()
print(len(word_features2))
print(word_features2[:50])

[ ] 3091
['aakash', 'aatif', 'ab', 'abd', 'abdevilliers', 'abhishek', 'able', 'absence', 'absorb', 'accept', 'acceptable', 'access', 'accomplishment', 'according', 'acc

[ ] vectorizer3 = TfidfVectorizer()
X3 = vectorizer3.fit_transform(filtered_headlines.values.astype('U'))
words = vectorizer3.get_feature_names()

[ ] <25x3091 sparse matrix of type '<class 'numpy.float64''
with 5766 stored elements in Compressed Sparse Row format>

[ ] from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters = 10, n_init = 20, n_jobs = 1)
kmeans.fit(X3)
# We look at 5 the clusters generated by k-means.
```

Figure 5.5: Module to Train K-Means Model

Fig. 5.5 shows that this module is trained with k-means algorithm. Library *Sklearn* is used to train the k-means algorithm. The *Tfidfveterizer* is used to get the word features and after tokenization the meaningful words called tokens are identified. The main use of tokenization is to split the sentences into individual tokens. For example ‘There are readers who prefer learning’ so in this sentence ‘there’, are, ’readers’, ’who’, ’prefer’ and learning’ are the tokens. Then using sklearn library k-means algorithm has imported to make the cluster based on the results on the bing_data CSV file.

Step 3: The collection of documents contains some unnecessary words due to which dimensionality of document increases. Pronoun, adverb, preposition etc. which are used throughout in the document has to be removed to get proper result. For example ‘Can listening be exhausting’ so after removal of stop word can and be it will results in Listening, exhausting. The below screen shot show that the different cluster heads from various clusters have been stored in array.

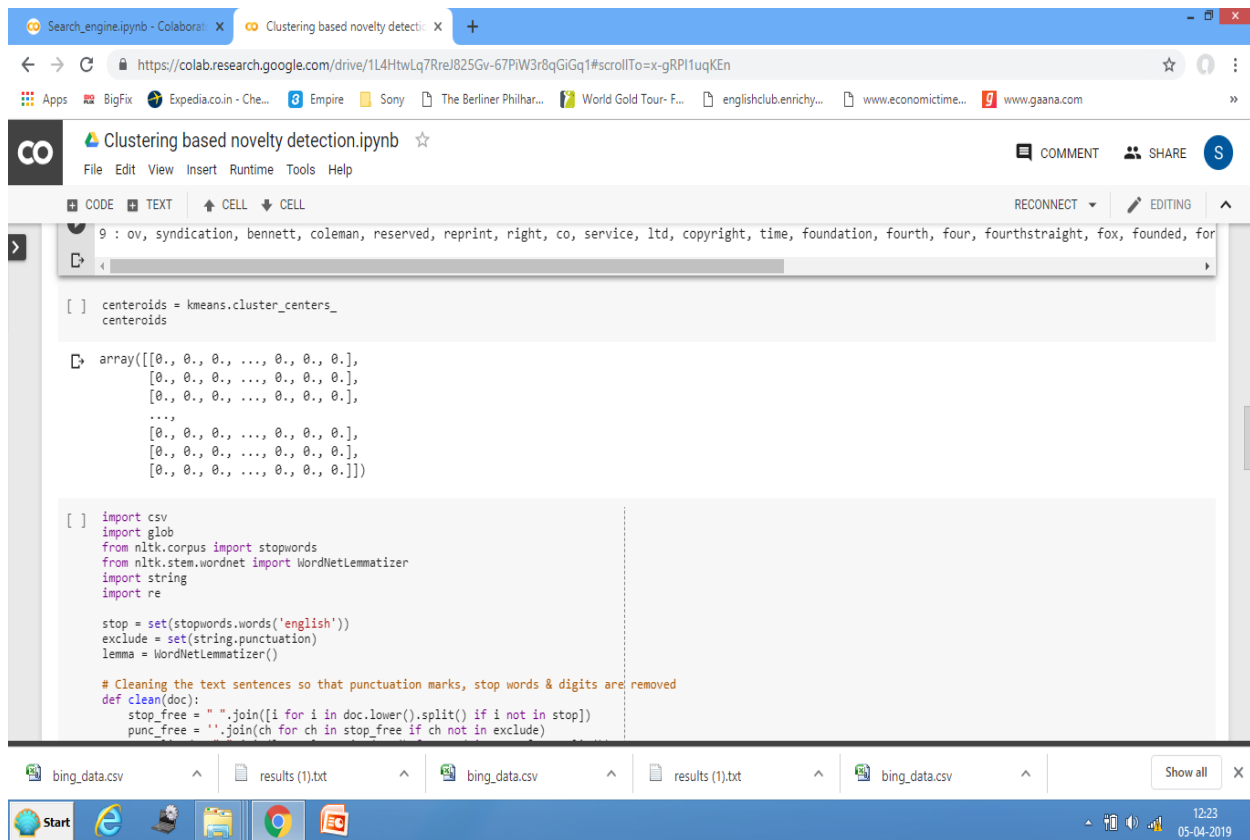


Figure 5.6: Array for Cluster head

Step 4: In this step the novel documents from thirty documents have been extracted and store in file on the disk.. The novel results have been generated by selecting the centriod with minimum distance from each cluster. The centriod from each cluster has large distance from the other centriod. Therefore, these documents found as novel.

5.5 RESULT ANALYSIS

Table 5.3 shows that different queries are fired in different domains on the Bing Search Engine Interface.As in the table thirty documents retrieved corresponding to the query ‘holi’, IPL, Narendra Modi and Pulwama attack. Result analysis shown below when manually find the novel documents based on result retrieved by Bing search engine and proposed approach:

Table 5.3: Comparison of Bing search engine with Proposed Approach

Domain for Query/keyword	Bing Search Engine Retrieved Documents	Bing Search Engine		Proposed Approach	
		Novel documents out of 30 documents	Novel documents in first ten documents	Novel documents out of 30 documents	Novel documents in first ten documents
Festival (Holi)	30	09	05	09	09
Sports (IPL)	30	07	02	08	08
Politics(Narender Modi)	30	08	05	09	09
Police force(Pulwama Attack)	30	09	05	09	09

Result Analysis 1: As shown in the table 5.3, figure 5.7 and figure 5.8 that Bing search engine give 09 novel documents from 30 documents for query 'holi' and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.

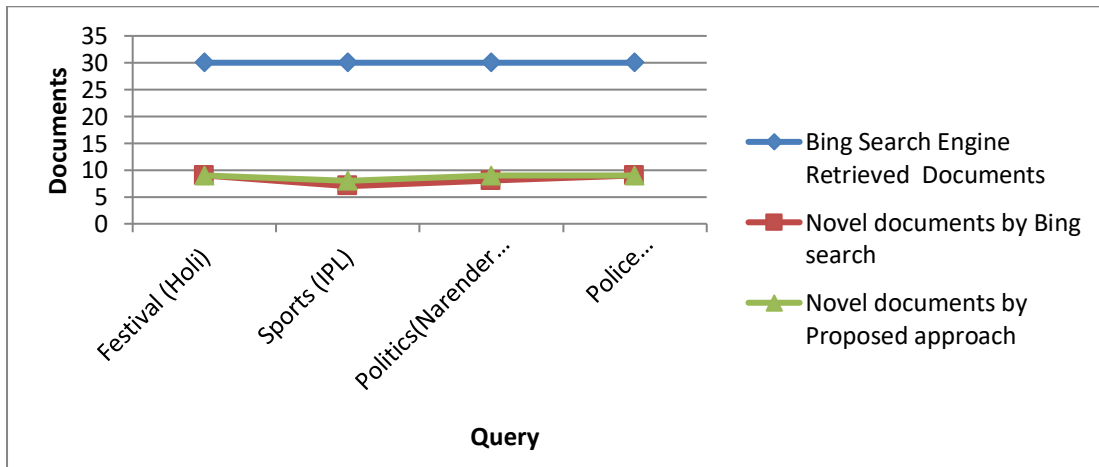


Figure 5.7: Novel documents out of 30 documents

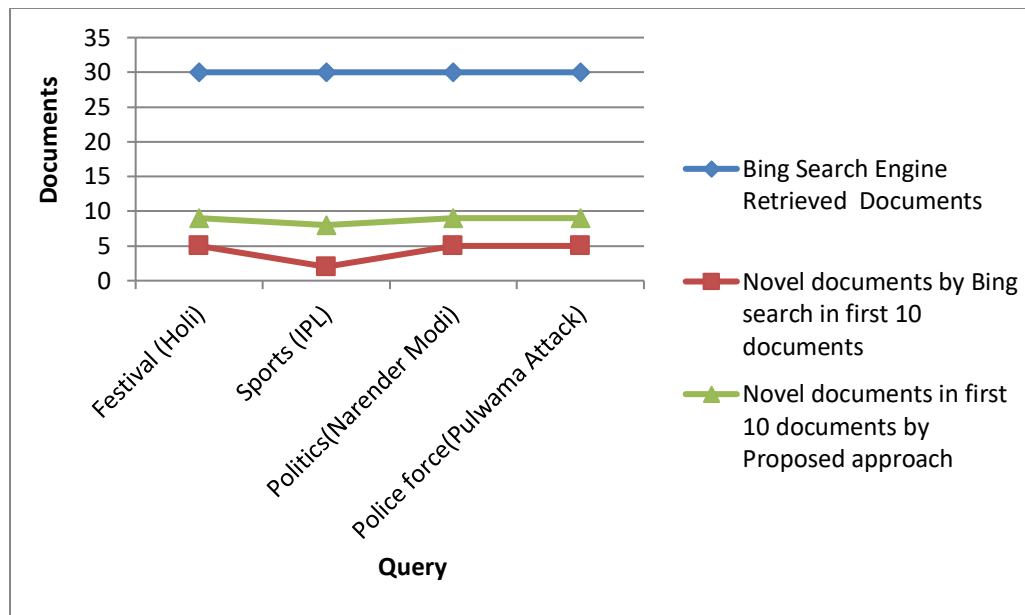


Figure 5.8: Novel documents out of the first 10 documents

Result Analysis 2: As shown in the table 5.3, figure 5.7 and figure 5.8 above that Bing search engine give 07 novel documents from 30 documents for query 'IPL' and only 02 novel documents from the first 10 documents. On the other hand proposed approach give 08 documents which all are novel and filter out the remaining 22 documents out of 30.

Result Analysis 3: As shown in the table 5.3, figure 5.7 and figure 5.8 above that Bing search engine give 08 novel documents from 30 documents for query 'narendra modi' and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.

Result Analysis 4: As shown in the table 5.3, figure 5.7 and figure 5.8 above that Bing search engine give 09 novel documents from 30 documents for query 'pulwama attack' and only 05 novel documents from the first 10 documents. On the other hand proposed approach give 09 documents which all are novel and filter out the remaining 21 documents out of 30.

5.5.1 Performance Evaluation

The proposed clustering based approach is compared with Bing search engine in terms of Precision, Recall and F-Score are defined as:

$$\text{Precision } (P) = \frac{\text{Novel documents detected}}{\text{Total number of retrieved documents}} \quad (5.2)$$

$$\text{Recall } (R) = \frac{\text{Novel documents detected}}{\text{Total Novel documents in the data set}} \quad (5.3)$$

$$F - \text{Score} = \frac{(2 * P * R)}{(P + R)} \quad (5.4)$$

Table 5.4 lists out the precision, recall and F-score values for keyword similarity for general document retrieval and proposed technique:

Table 5.4: Precision, Recall and F-Score Comparison

Domain for Query/keyword	Count of Documents Considered	Bing Search Engine Retrieved Documents			Proposed Approach		
		Precision	Recall	F-Score	Precision	Recall	F-Score
Festival (Holi)	30	0.16	0.55	0.24	0.30	0.90	0.45
Sports (IPL)	30	0.06	0.28	0.09	0.26	0.80	0.39
Politics(Narender Modi)	30	0.16	0.62	0.25	0.30	0.90	0.45
Police force(Pulwama Attack)	30	0.16	0.55	0.13	0.30	0.90	0.45

The precision, Recall and F-Score plots are shown in Figures 5.9, 5.10 and 5.11 respectively

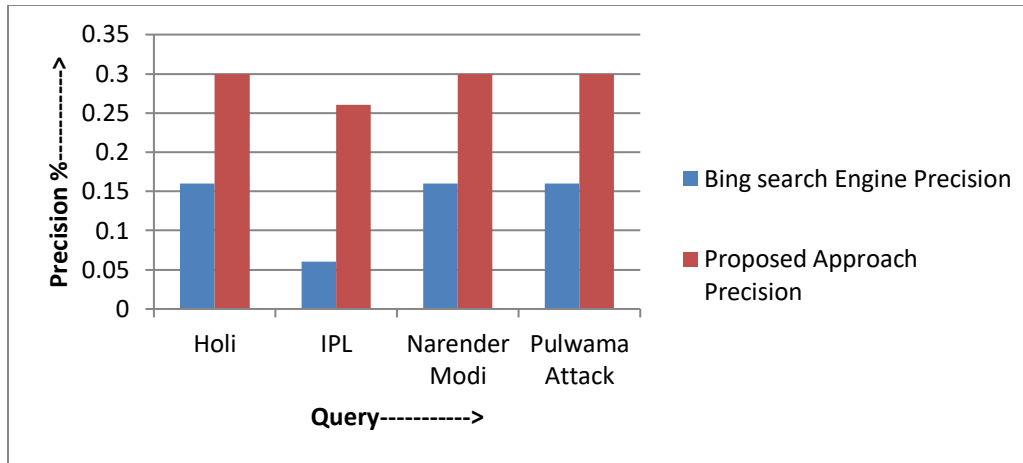


Figure 5.9: Precision Plot

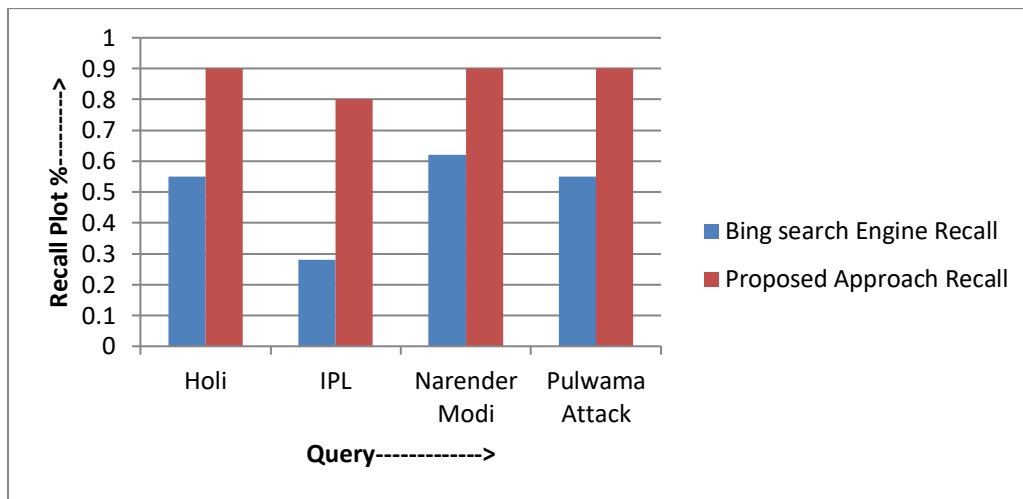


Figure 5.10: Recall Plot

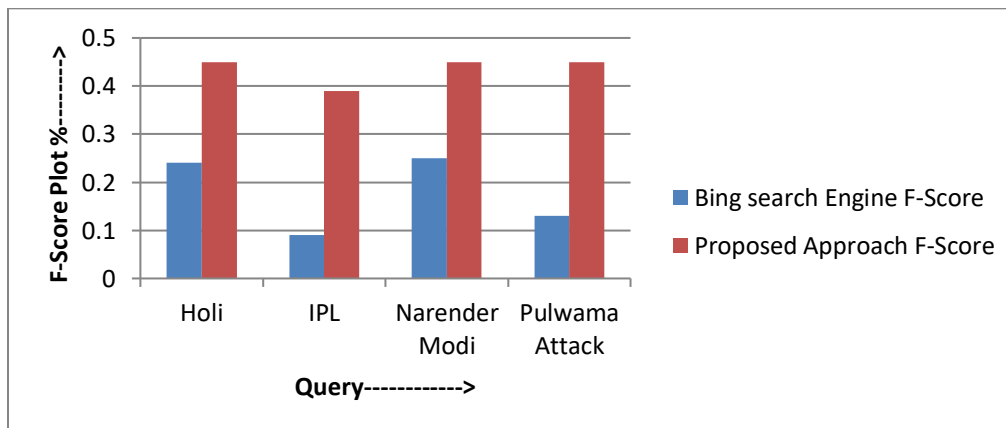


Figure 5.11: F-Score Plot

From the implementation it has been proved that this proposed method gives proper result in lesser amount of time and with better efficiency. The precision, recall and F-measure are also high when this algorithm applied on the set of documents as compared with Bing Search Engine. So that the proposed approach provides the novel documents based on the given query and filter out the redundant documents

5.6 SUMMARY

In this work, clustering based approach for novelty detection has been investigated and tested on the set of documents. The incoming stream of documents based on the query has been clustered using k-means clustering algorithm and then the clusters head are calculated. The cluster heads selected from different clusters retrieved novel documents. The work has been compared the proposed approach with the results given by Bing Search Engine according to the query in different domains. The proposed approach provides the novel documents based on the given query and filter out the redundant documents

CHAPTER VI

SEMANTIC SIMILARITY AND TEXT SUMMARIZATION BASED NOVELTY DETECTION

6.1 INTRODUCTION

Flow web crawlers search the questions at extremely fast, however the issue of curiosity location or repetitive data actually endures. It burns-through valuable time and memory of clients looking for the new record over the web.

In this chapter, an innovative novelty detection mechanism is proposed, which can be appended with the current web crawlers. The proposed mechanism first summarizes the text, based on ontology [72,73], and then from the obtained summary [74], semantic similarity [79] is calculated using word net 3.0. The hash value is then calculated using the winnowing [69] algorithm. This hash value of the document is matched with others using the Dice coefficient to calculate the similarity index. Based on the threshold chosen for similarity, the document is treated either as novel or not. This proposed mechanism is implemented using SQL as backend and visual studio-2012 as frontend.

The problems associated with the current search engine and crawlers [78] are as listed below:

- One issue with a focused crawler [78] is that they miss essential pages by just creeping pages that are relied upon to give immediate benefits.
- The crawlers download numerous unimportant pages that lead to the utilization of system transfer speed. They receive pooling technique for the upkeep of freshness of database.
- The collaborative crawler [78] utilizes the gathering of crawling nodes; it is conceivable that several nodes download a similar page many times. Therefore needs to build up a method that decreases these cover of pages.
- The parallel crawler [78] having many crawling methods called C-Proc's. At the point when various C-Proc's are working freely, it is conceivable that more than one may download a similar page on many times.

To resolve the above said setbacks this chapter introduces a novelty detection technique which overcomes the concern of downloading these redundant pages.

6.2 GENERIC CRAWLER METHODOLOGY

In this work, firstly, a generic crawler is proposed that takes a query on a specific domain, and the crawler results are stored in an indexed database. The database stores the URL of the query together with its HTML tags, metadata tags, etc. The URL enter by the admin are stored in the dictionary. This method also provides a search interface on which the user can apply a query based on a specific domain, stored in the database. When the user types a keyword on the search interface, it shows the web pages stored in the database. The retrieved list of web pages may contain relevant and redundant results, which is a time-consuming task for the user to read all the pages. The architecture of generic crawler shown in figure 6.1:

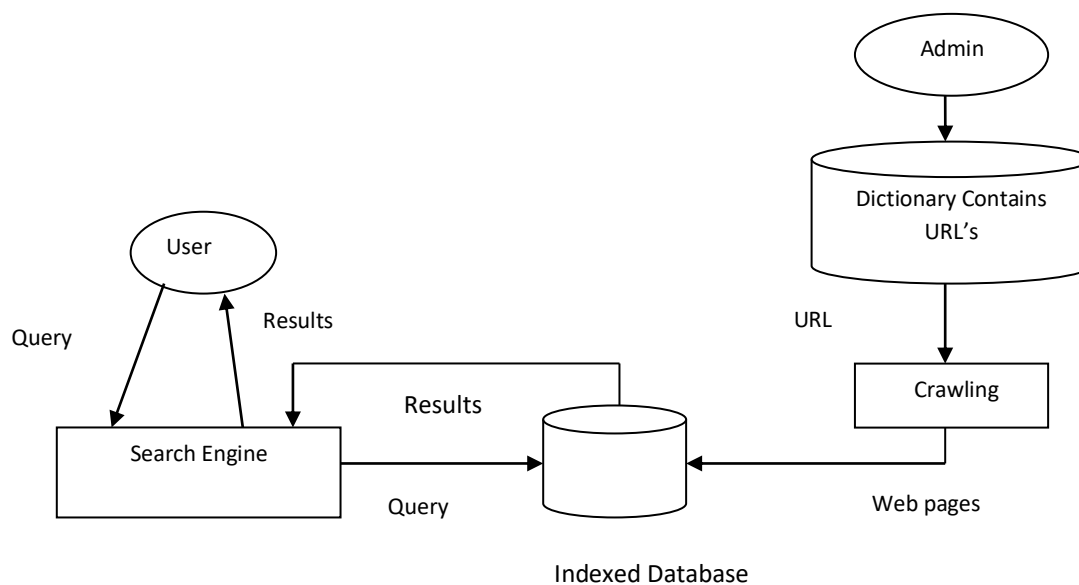


Figure 6.1: Generic Web Crawler Architecture

Fig. 6.2 shows the interface for domain-specific Generic Crawler, which includes website category related to education, politics, sports, technology, health, entertainment, travel, and zoology. Users can enter the website related topic and website link, as shown in the interface.

Then the user clicks on the crawl button, and the generic crawler crawls the web pages based on the website URL (Uniform Resource Locator) to store them into the database.

Base Crawler - Novelty

Web Site Category* Others
Name of Web Site* code project
Web Site* http://www.codeproject.com Crawl

Fields marked * are mandatory.

[search](#)

Figure 6.2: Generic Crawler Novelty Interface

Figure. 6.3 shows the SQL database, which includes three tables T_Category, T_website, and T_webpages. The table T_Category includes the website categories i.e., education, politics, sports, technology, health, entertainment, travel, and zoology. The T_Website and T_Webpages store the website related information together with web pages related information. Upon executing the query by select command, the information from the database table is shown in the screenshot. The database stores the URL of the query together with its HTML tags, metadata tags, etc.

The indexed database contains the following data set from three different domains:

- **Data set 1:** This data set consists of 1634 documents from domain like sports, politics, technology and education.
- **Data set 2:** This data set consists of 4430 documents from domain like health, entrainment, travel and zoology.
- **Data set 3:** This data set consists of 4385 documents from domain like science, business, world and transport.

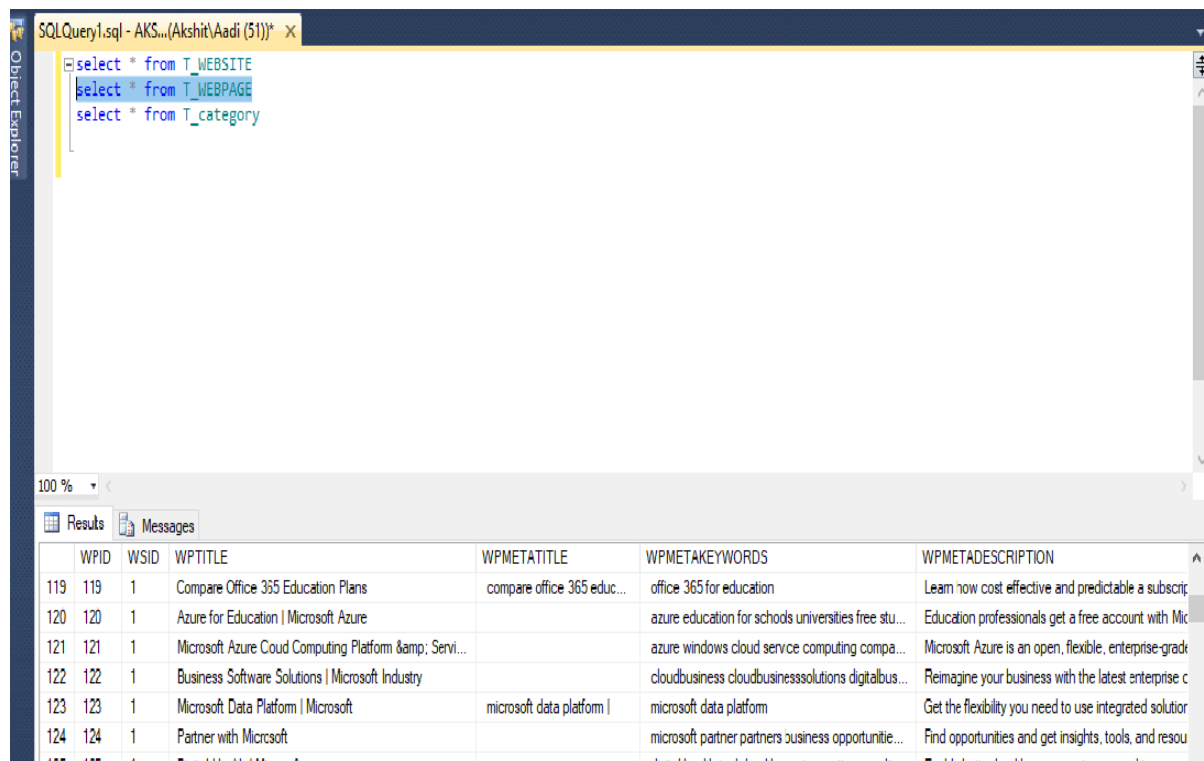


Figure 6.3: SQL Indexed Database

6.3 PROPOSED CRAWLER METHODOLOGY FOR NOVELTY DETECTION

In the proposed methodology, the limitation of a generic crawler that is repeated occurrence of the redundant documents is eliminated. The projected method provides the relevant and novel results to the user and filters out the redundant ones. This work includes extractive text summarization using ontology to calculate the summary of the text document after that the Winnowing fingerprint algorithm [93] is applied for similarity calculation and Word Net 3.0 [76] for semantic similarity [75,76]. Winnowing calculation is a technique for word comparability search in a document by looking at the fingerprint on the document. The algorithm input is the text document, which is processed and yield as a hash value. The hash value is then called as the unique finger impression, which is utilized to look at the comparability of each document. The difference of Winnowing Calculation and another calculation of comparability indicator is in the choice procedure of its fingerprint. The consequence of hash value figuring is partitioned into window w in which the smallest value will be taken from every window for the document fingerprint. The stepwise procedure for proposed mechanism is given below:

- Initially the text summarization technique is applied using ontology, which provides the relevant sentences.
- Assume the target text and the original text are as strings with the length t .
- N-grams [80, 81] are generated from tokens to obtain documents with fixed-length strings.
- N-grams [82] are further processed for discovering hashes, which are collected in order to diminish the size of documents.
- The strings are reformed to some numeric values called hashes [93]. A suitable similarity measure is applied to hashes for similarity determination.

The architecture of the proposed methodology is shown in figure 6.4:

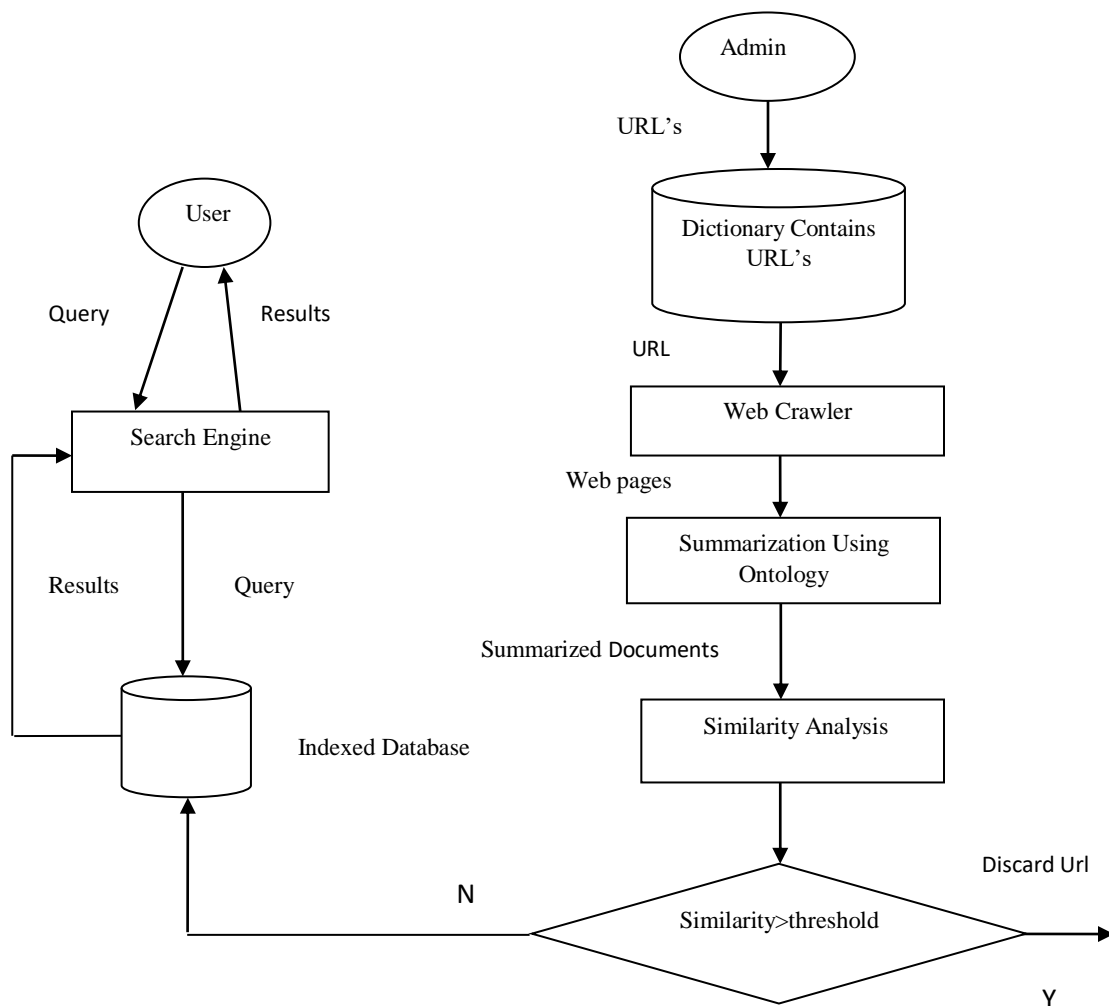


Figure 6.4: Proposed Architecture

6.3.1. Algorithm for the Proposed Crawler Novelty Detection

Input: URL (source code in S_{out}), DB \rightarrow Data Base

DB first row say in $S_{current}$

Begin

Step 1: Fetch source code in S_{out}

Step 2: Summarization of fetched data S_{out}

Step 3: for each row in DB

3.1 Summarization of each row of DB ($S_{current}$)

Step 4: Find the similarity of both summarized data (S_{out} , $S_{current}$)

Step 5: if (similarity > threshold)

5.1 Break the loop and will not compare with any row

5.2 Because it already finds a similar row

Else Check with another DB row

Step 6: If it does not find the similar doc in DB, than save a new row i.e. fetched data in DB.

End.

6.3.2. Detailed steps for Proposed Crawler Novelty

In figure 6.5, steps for the ontology-based text summarization are shown. It consists of sentence parsing, tokenization, and stop word removal, noun filtering using WorldNet 3.0, word overlap calculation and minimization of summarized data using ontology. These steps are explained in brief as below:

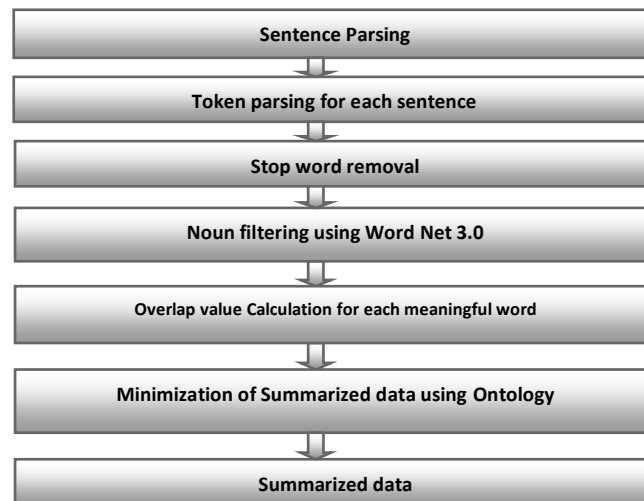


Figure 6.5 Ontology based Text Summarization

6.3.2.1 Sentence Parsing

A document comprises of countless sentences, so the archive is parts into sentences. The sentences comprise of an enormous number of words, and it isn't generally fundamental that each word is of significance. Because of which high dimensionality of the record must be decreased by handling the report to dispose of additional words to get the heaviness of every one of the word to be utilized in the calculation.

6.3.2.2 Tokenization

The cycles associated with data recovery require the expressions of documents. Tokenization is utilized to distinguish important words called tokens. The essential utilization of tokenization is to part the sentences into singular tokens. For instance 'There are perusers who favour learning' so in this sentence 'there', are, 'perusers', 'who', 'like' and learning' are the tokens.

6.3.2.3 Stop-words Removal

Stop words are much of the time happening, irrelevant words that show up in a data set record, article, or a page, and so forth pronoun, qualifier, relational word and so on which are utilized all through in the record must be taken out to get appropriate outcome. For instance 'Can listening be depleting' so after evacuation of stop word can and be it will brings about Listening, debilitating.

6.3.2.4 Noun Filtering using Word net 3.0

The sizeable lexical data set is utilized for English language to think about for word likeness. The rendition utilized in this examination is Word net 3.0 [76], which has 117,000 equivalent set. These equivalent set are called synset. Word net has a way connection between thing and action word action word as it were. This relationship is missing for other grammatical feature.

Example: A Tour is a long excursion on a boat or in a space apparatus

In the above sentence tour, excursion, boat, and space apparatus are nouns

6.3.2.5 Word Overlap Value Calculation

Subsequent to distinguishing the important words, the overlap [78] values are determined between the sentences in the report. Overlap value implies that how much comparative the words are in a sentence S1 and sentence S2. Essentially, the overlap is determined for sentence S3, etc. Amount of all the cover value speaks to the heaviness of the sentence and high value sentence is utilized in the summary. The determined overlap values are organized in diminishing order, and the initial three most noteworthy values have been remembered for the summary between two documents.

6.3.2.6 Minimization of Data Using Ontology

The Ontology is utilized to limit the summed up information further. Ontology of various domains for example Sport, Technology, Education and Politics, and so forth are put away in the information base table name ontology. Ontology gives a typical jargon of a zone and characterizes, at various degrees of convention, the importance of terms and connections between them. The connection between token1 to token2 in a specific sentence is given by with 'is a' connection. The tokens are additionally coordinated with sentence S1 to condemn S2 to additionally limiting the summed up information. Subsequently ontology tells about the significance of the terms in a specific sentence to additionally summarize the information. The last summarized information is gotten after this step on which similarity count to be performed.

6.4 SIMILARTY CALCULATION OF SUMMARIZED DATA

N-grams token-based MD5 function, Winnowing fingerprint matching algorithm using dice coefficient are used for similarity calculation of the summarized data. The steps are explained in brief as below:

6.4.1 N-Gram Formation

N-Gram [80,81], development is a cycle of changing over a string into substring. N is utilized for speaking to a number and tells the number of words will be picked in one gram. The contribution for N-Gram is a pre-handled string, and the estimation of N. N can be 1, 2,.....,n relying on the client's necessity. On the off chance that we take N=1, the N-Grams framed are known as unigrams. On the off chance that we take N = 2, at that point the grams framed are bigrams. In

the event that we take N=3, at that point grams shaped are trigrams, etc. N-grams are the continuous succession of N character cut of a string. They can be assessed utilizing $N = (p - m + 1)$, where p speaks to various letters in the archive and m speaks to the size of N-grams. N-grams are produced from tokens after evacuation of spaces as demonstrated as follows:

For size of N=5.

Let a string MynameIsHash, 5-Grams got from the string

Mynam yname nameI ameIs meIsH eIsHa IsHas sHash

6.4.2 Hash Conversion

An ASCII value speaks to each character of gram. It changes over grams into relating hash value [93]. Each character is changed over to ASCII value. Hashing is a cycle of transformation of grams into short fixed-length value. It is performed on the grounds that it is anything but difficult to track down short length value than to locate the original string. The hunt process will include and afterward utilizing it to discover a counterpart for a given value. Hashes are shaped to try not to overpower calculations. Therefore, require a piece of n grams to be utilized for examination, and the n grams are changed over to hash values.

The equation for hash formation can be given by

$$H(dk) = d1 * m^{(k-1)} + d2 * m^{(k-2)} + + d^{(k-1)} * m + d^k \quad (6.1)$$

In the equation 6.1, ASCII character is denoted by d, m denotes the basis of primes, and k represents the value of k-grams.

The contribution to hash function can be self-assertive, however the output is fixed alluded to as n bit. This cycle is the hashing of information. The changed over little values are the hexadecimal value to be changed over into decimal. A few hashes are made for each archive relating to every n-gram of the document.

6.4.3 Frame Parsing

It is a method of converting hashes into frames. The input provided contains two parameters. The first parameter is the hash value, which is the output of the MD5 method. The second parameter is n , which tells the number of hashes to be kept in the frame. In this work, we have taken n as '4'. This parsing is done to ensure that minimum value is always available for selection from each frame. A function substring is used for providing the value of ' n '. The frames are created according to the size of n . The output is stored in an array list. Each frame will contain a value that will be used for comparison of two documents. Each frame will contain an equal number of hashes in it.

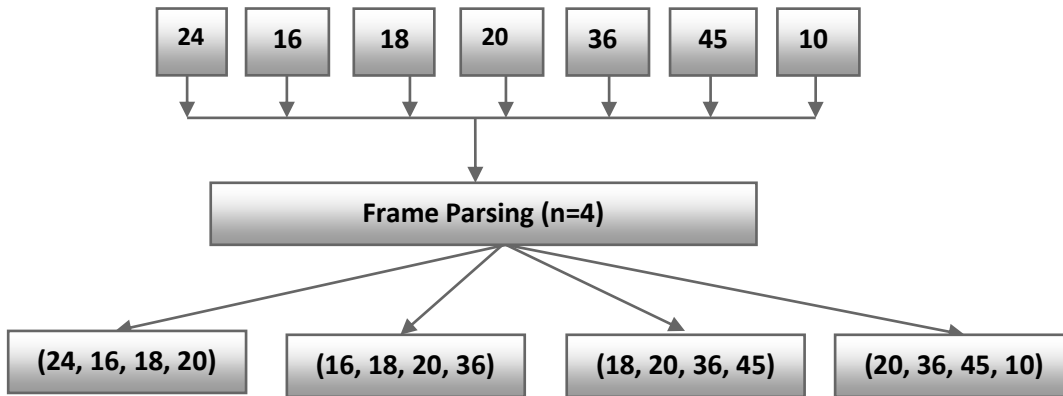


Figure 6.6: Frame Parsing

6.4.4 Process of Fingerprint Selection

In the previous phase, frames of equal size were formed. Each frame contains an equal number of hashes. For further processing of data, we need to choose a minimum value from each frame. All the values in a frame are compared to each other to find the least value. The reason for choosing the value as a minimum is that, the least value in one frame is likely to be the least value in other frames too. It said that a minimum of ' n ' random number is smaller than one additional random number. The number of the values selected is much less as compared to the number of frames. It makes the document be represented by a small number of values and provides scalability to documents. When there are two similar hash values in two frames, then the value in the rightmost frame is

chosen to be the least hash value. These all selected hash values together represent a document. This process uses the looping function to select the values on each window and array function to ensure that there is no similar value on the array as the result of fingerprint selection. The least hash values selected from figure 6.6 are 10 and 16.

6.5 CALCULATION PROCESS OF DOCUMENT SIMILARITY

The Dice Coefficient of two sets is a proportion of their convergence and it is scaled by their Size (giving the incentive in the reach 0 to 1). It is determined as convergence over association of qualities.

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|} \quad (6.2)$$

Let us take a string likeness measure the coefficient can be determined for two strings X and Y as follow

X= night and Y= nauht, we would find the set of bigrams in each word

{ni, ig, gh, ht} and {na, au, uh, ht}

Each set has four elements, and the intersection of these two sets has only one element ht. Inserting, these numbers into the formula, we calculate

$$Similarity = \frac{2 * 1}{4 + 4} = 0.25 \quad (6.3)$$

This Similarity is compared with a pre-defined threshold if it is greater than the threshold it provides that the current web page is similar to the pages already in the database. This page will not add to the database.

6.6 SIMULATION SET PARAMETERS

The proposed algorithm used similarity calculation that tells whether the new web page added into the database depends on a threshold value. By experimenting with similarity values calculation, the simulation setup parameter threshold set to be 0.65. If the similarity index is higher than this value, the web page does not add to the database because it already exists their; otherwise, it will be compared with other rows in the database. The page that is added to the database would be novel to other pages or documents into the database. In this way, the database will store only the novel pages at the crawling time, and search results always provide the novel results to the user query.

6.6.1 Set up parameters

The setup parameters are hardware, software and dataset used to perform the experiments. Table 6.1 show the setup parameters used to develop the overall experiments.

Table 6.1: Setup Parameters

Processor	Intel i3 , 1.90 GHz processor	
Memory	RAM 4.00 GB, HDD 500 GB	
Software	Windows 10 Operating System, Microsoft Visual Studio 2012 (.NET) as front end, SQL server 2012 as a back end database.	
Data Set 1 (1634 documents)	Domain set 1	Sports
		Politics
		Education
		Technology
Data Set 2 (4430 documents)	Domain set 2	Health
		Entertainment
		Travel
		Zoology

Data Set 3 (4385 documents)	Domain set 3	Science
		World
		Business
		Transport

6.6.2 Performance Parameters

To measure the efficacy of the proposed scheme several performance metrics are taken given under:

- **Redundancy Removal (RR):** It is calculated as the difference between number of pages retrieved by the generic approach and number of pages retrieved by proposed approach.

$$RR = abs(NP_{GA} - NP_{PA}) \quad (6.4)$$

Where, RR is the Redundancy Removal, NP_{GA} is the number of pages retrieved by the generic approach, and NP_{PA} is the number of pages retrieved by the proposed approach

- **Memory Overhead(MO):** For calculations of memory overhead number of pages retrieved by the generic approach and number of pages retrieved by proposed approach multiply by the page size are computed. This is the memory overhead used and is given by

$$MO = NP_R * P_S \quad (6.5)$$

Where, MO is the Memory overhead, and P_S is the page size in Megabytes.

- **Number of Pages Identified (NPI):** This gives the number of pages identified after the given search, which are relevant and novel. This is given as

$$NPI = NP_{PA} \quad (6.6)$$

Where, NPI , is the number of pages identified by the proposed approach, which is same as the pages retrieved by the proposed approach

6.7 IMPLEMENTATION

The implementation includes the Microsoft Visual Studio 2012 (.NET) as a front end and SQL server 2012 as a back end database. The SQL database includes three tables T_Category, T_website, and T_webpages. The table T_Category includes the website categories i.e., education, politics, sports, technology, health, entertainment, travel, and zoology. The T_Website and T_Webpages store the website related information together with web pages related information. The database stores the URL of the query together with its HTML tags, metadata tags, etc. It also includes the table ontology, Senti Dictionary table (Word Net 3.0), and overlap table to store ontology together with overlap calculation information. The dictionary contains the URLs used by the user to search any query word.

Figure 6.7 is a Search Engine interface that appears when the user clicks on the search button. This interface includes keyword to be search based on the advanced search and field-specific search.

The screenshot displays a web-based search engine interface. At the top, the title "Search Engine" is prominently displayed. Below the title is a search bar containing the text "swayam" and a "Search" button. Underneath the search bar, there is an "Advanced Search" section. This section includes a "No. of Results Per Page" dropdown menu set to "5" and a "Language Selection" dropdown menu with "English" selected. Below the advanced search section is a "Field Specific Search" section. This section contains a grid of checkboxes for various categories: Education (checked), Travel, Politics, World, Health, Science, Sports, Transport, Business, and Others (checked). At the bottom of this section, there are links for "Select None" and "Select All".

Figure 6.7: Search Engine Interface

Figure 6.8 showed the search results when the user typed the keyword or topic on the search interface to the generic crawler. It shows the results for the keyword 'code,' which already stored

in the database for the technology category. This search result is showing the redundant and relevant webpage for the given query, which is a tedious and time-consuming task for the user to read whole documents. The proposed methodology included the text summarization, syntactic similarity, plus semantic similarity to overcome the limitations of the generic crawler. This work provides the relevant and novel results to the user's query and filters out the redundant ones.

Search Results

For code

Search Again

Total Records: 344

[free source code tutorials](#)

Free source code and tutorials for Software developers and Architects.; Updated: 6 May 2019

<http://www.codeproject.com>

[free source code tutorials](#)

Free source code and tutorials for Software developers and Architects.; Updated: 6 May 2019

<https://www.codeproject.com>

<https://www.codeproject.com/webservices/LoungeRSS.aspx>

[free source code tutorials](#)

Free source code and tutorials for Software developers and Architects.; Updated: 6 May 2019

<http://www.codeproject.com/>

[free source code tutorials](#)

Figure 6.8: List of WebPages for the Query ‘code’ on Generic Crawler Search Interface

When the same query as in figure 6.8 runs on the proposed crawler interface as in Figure 6.9, it filters out redundant ones and displays only the relevant and novel pages.

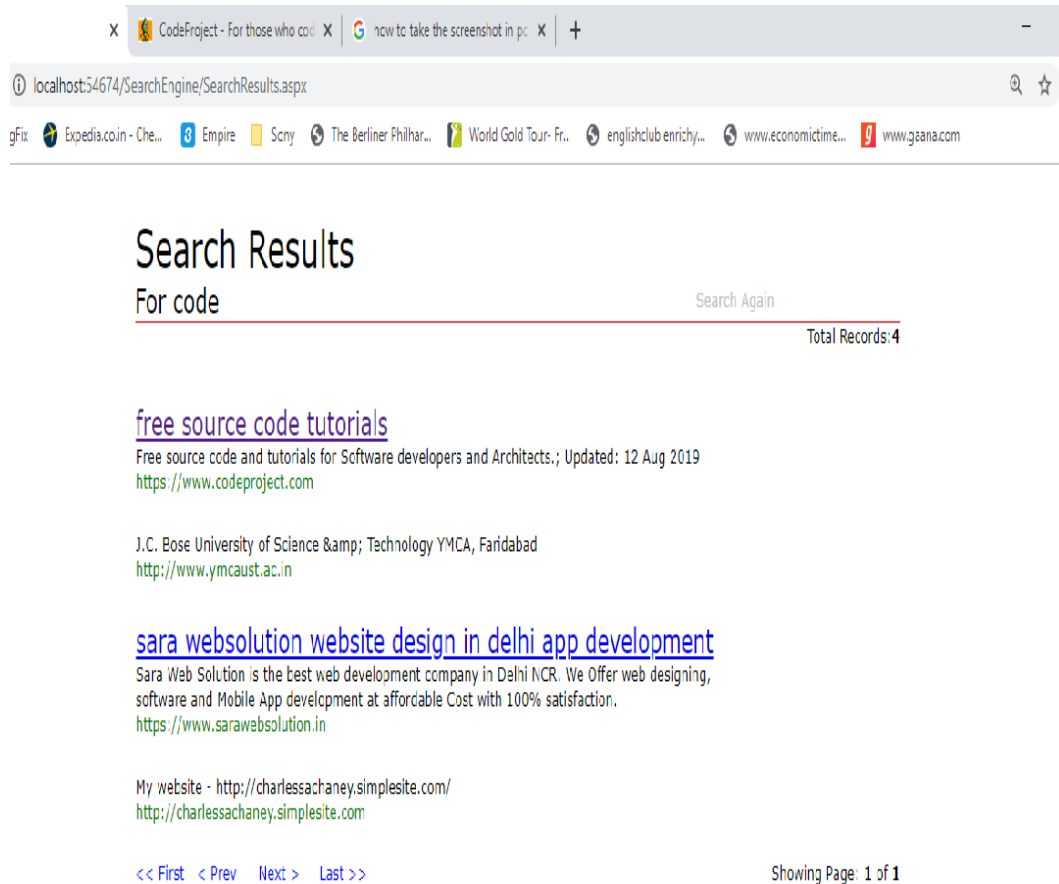


Figure 6.9: List of WebPages for the Query ‘code’ on Proposed Crawler Search Interface

6.8. RESULTS AND DISCUSSION

Users can press on the search button after typing any topic or keyword on the search interface with the number of pages to be displayed together with ticking on the field-specific search i.e. education, sports, and politics, etc.

6.8.1. Data Set 1: Table 6.2 shows the different queries that executed for different domains on the Search Engine Interface of the generic crawler and proposed crawler novelty. The results of these queries are stored in the crawler indexed database of the generic method and proposed method.

Table 6.2: Comparison of Generic Crawler and Proposed Crawler Novelty

Domain	Query	Generic Crawler (No. of Pages Retrieved)	Proposed Crawler (No. of Novel Pages Retrieved)	Redundant Pages
Sports	Sports	80	5	75
	Ball	70	3	72
	Boxing	40	2	38
	Cycling	30	1	29
Politics	Politics	100	4	96
	Election	60	3	57
	Campaign	40	2	38
	Leadership	40	2	38
Education	Education	170	9	161
	YMCA	50	1	49
	University	110	5	105
	Board	140	6	132
Technology	Code	344	4	340
	Web	130	8	122
	HTML	130	8	122
	Java	120	6	114
Total		1654	69	1588

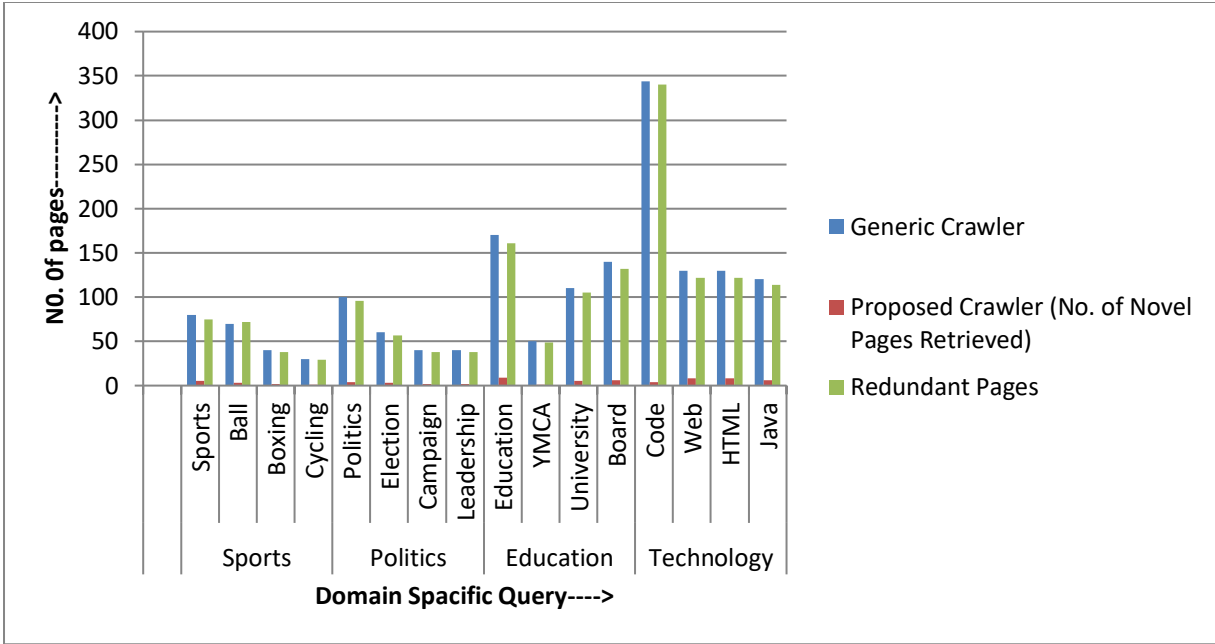


Figure 6.10: Comparison of Generic Crawler and Proposed Crawler Novelty Results

Result Analysis 1: As shown in table 6.2 and figure 6.10 above that Generic crawler search provide 80,70,40 and 30 documents for queries '*sports*', '*Ball*,' '*boxing*,' and '*cycling*' under the domain '*Sports*', respectively. On the other hand, the proposed approach provides 5, 3, 2, and 1, which all are novel and filter out the remaining ones.

Result Analysis 2: As shown in table 6.2 and figure 6.10 above that Generic crawler search provide 100,60,40 and 40 documents for queries '*politics*', '*election*,' '*campaign*,' and '*leadership*' under the domain '*Politics*', respectively. On the other hand, the proposed approach provides 4, 3, 2, and 2, which all are novel and filter out the remaining ones.

Result Analysis 3: As shown in table 6.2 and figure 6.10 above that Generic crawler search provide 170,110,50 and 140 documents for queries '*education*', '*ymca*,' '*university*,' and '*board*' under the domain '*Education*', respectively. On the other hand, the proposed approach provides 9, 1, 5, and 6, which all are novel and filter out the remaining ones.

Result Analysis 4: As shown in table 6.2 and figure 6.10 above that Generic crawler search provide 344,130,130 and 120 documents for queries '*code*', '*web*,' '*html*,' and '*java*' under the domain '*Technology*', respectively. On the other hand, the proposed approach provides 4, 8, 8, and 6, which all are novel and filter out the remaining ones.

Memory Overhead: As shown in figure 6.11, If a page of size is 5MB, then the generic approach memory requirement is $1654 \times 5 = 8270$ MB, and according to the proposed approach, the memory requirement is $69 \times 5 = 345$ MB.

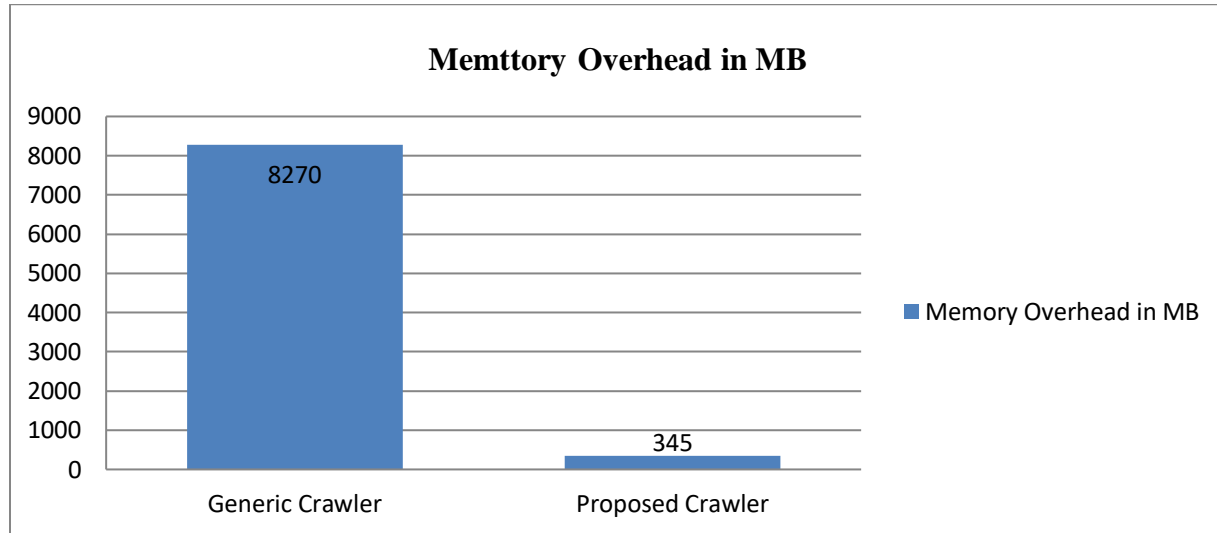


Figure 6.11: Comparison of Memory Overhead

6.8.2. Data Set 2: Table 6.3 shows the different queries that executed for different domains on the Search Engine Interface of the generic crawler and proposed crawler novelty. The results of these queries are stored in the crawler indexed database of the generic method and proposed method.

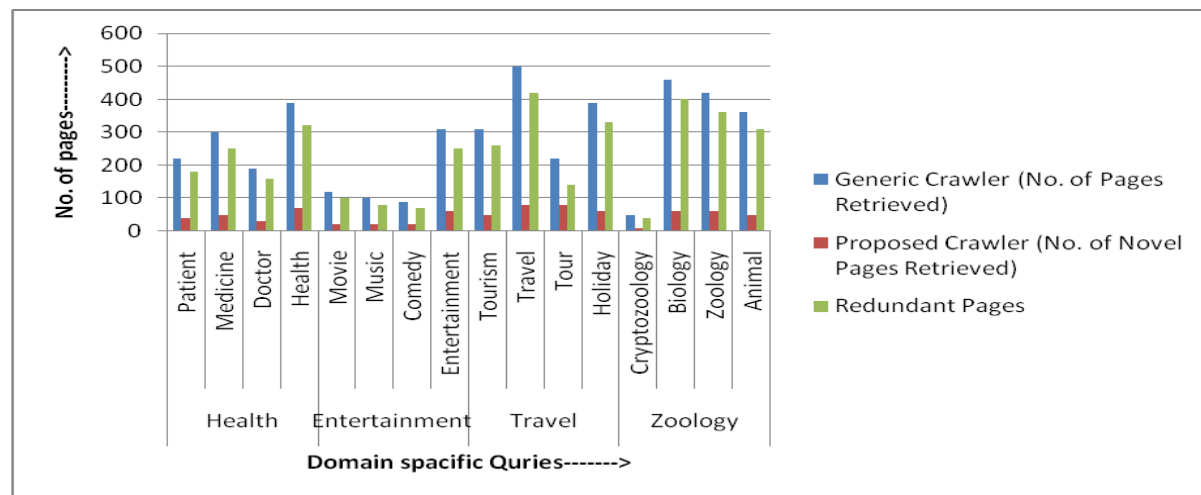


Figure 6.12: Comparison of Generic Crawler and Proposed Crawler Novelty Results

Table 6.3: Comparison of Generic Crawler and Proposed Crawler Novelty

Domain	Query	Generic Crawler (No. of Pages Retrieved)	Proposed Crawler (No. of Novel Pages Retrieved)	Redundant Pages
Health	Patient	220	40	180
	Medicine	300	50	250
	Doctor	190	30	160
	Health	390	70	320
Entertainment	Movie	120	20	100
	Music	100	20	80
	Comedy	90	20	70
	Entertainment	310	60	250
Travel	Tourism	310	50	260
	Travel	500	80	420
	Tour	220	80	140
	Holiday	390	60	330
Zoology	Cryptozoology	50	10	40
	Biology	460	60	400
	Zoology	420	60	360
	Animal	360	50	310
Total		4430	760	3670

Result Analysis 1: As shown in table 6.3 and figure 6.12 above, that Generic crawler search provides 220,300,190 and 390 documents for queries' patient', 'medicine', 'doctor,' and 'health'

under the domain name 'Health', respectively. On the other hand, the proposed approach provides 40, 50, 30, and 70, which all are novel and filter out the redundant pages.

Result Analysis 2: As shown in table 6.3 and figure 6.12 above that Generic crawler search provide 120,100,90 and 310 documents for queries' *movie*', *'music'*, *'comedy'*, and *'entertainment'* under the domain name *'Entertainment'*, respectively. On the other hand proposed approach provide 20, 20, 20 and 60 which all are novel and filter out the redundant pages.

Result Analysis 3: As shown in table 6.3 and figure 6.12 above that Generic crawler search provide 310,220,500 and 390 documents for queries' *tourism*', *'travel'*, *'tour'*, and *'holiday'* under the domain name *'Travel'*, respectively. On the other hand, the proposed approach provides 50, 80, 80, and 60, which all are novel and filter out the redundant pages.

Result Analysis 4: As shown in table 6.3 and figure 6.12 above that Generic crawler search provide 50,460,420 and 360 documents for queries' *cryptozoology*', *'biology'*, *'zoology'*, and *'animal'* under the domain name *'Zoology'*, respectively. On the other hand, the proposed approach provides 10, 60, 60, and 50, which all are novel and filter out the redundant pages.

Memory Overhead: As shown in figure 6.13, if a page of size is 5 MB, then the generic approach memory overhead is $4430 \times 5 = 22150$ MB, and according to the proposed approach, the memory overhead is $760 \times 5 = 3800$ MB.

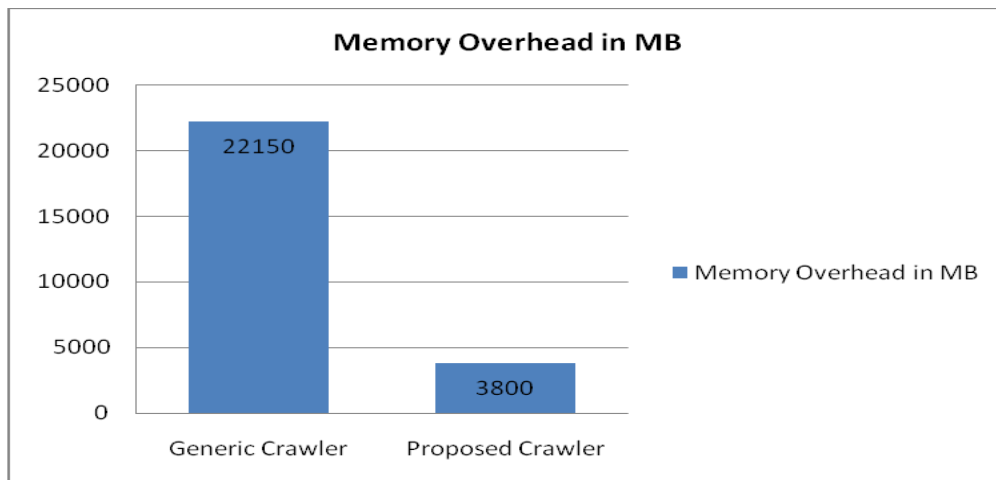


Figure 6.13: Comparison of Memory Overhead

6.8.3. Data Set 3: Table 6.4 shows the different queries that executed for different domains on the Search Engine Interface of the generic crawler and proposed crawler novelty. The results of these queries are stored in the crawler indexed database of the generic method and proposed method.

Table 6.4: Comparison of Generic Crawler and Proposed Crawler Novelty

Domain	Query	Generic Crawler (No. of Pages Retrieved)	Proposed Crawler (No. of Novel Pages Retrieved)	Redundant Pages
Science	Geophysics	130	15	115
	Scientist	100	20	80
	Laboratory	80	10	70
	Laws	160	22	138
World	Universe	390	45	345
	Nature	360	38	322
	Society	310	20	290
	People	320	28	292
Business	Service	280	39	241
	Merchandise	220	24	196
	Manufacturing	180	30	150
	Partnership	170	25	145
Transport	Bus	480	50	430
	Car	430	40	390
	Truck	415	30	385
	Vehicle	360	42	318
Total		4385	478	3907

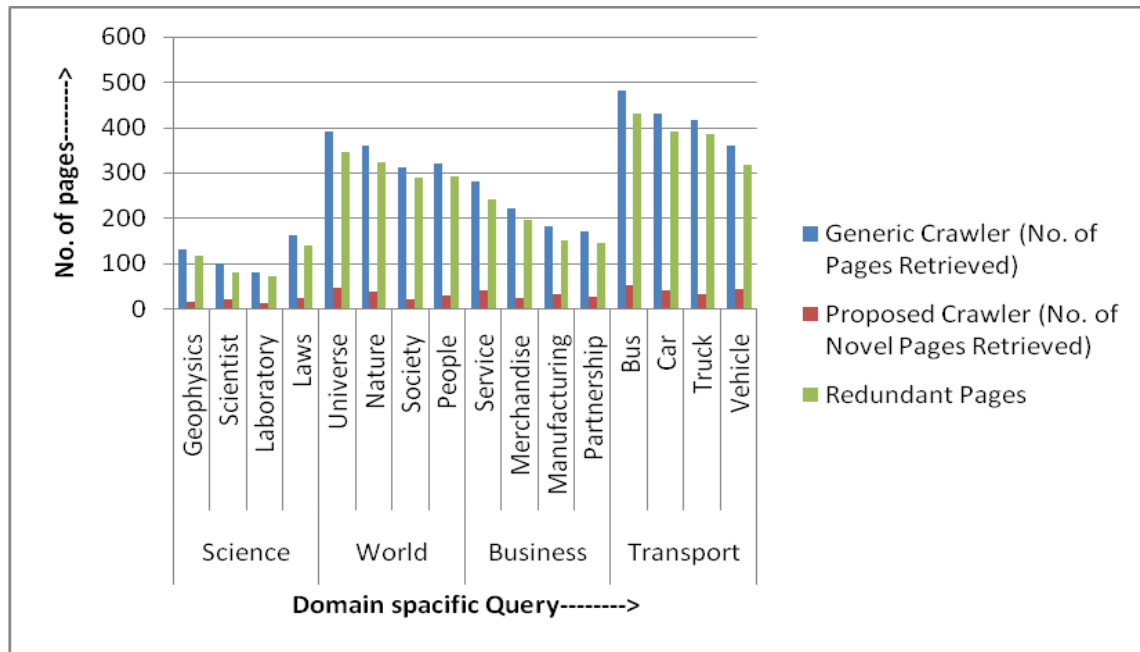


Figure 6.14: Comparison of Generic Crawler and Proposed Crawler Novelty Results

Result Analysis 1: As shown in table 6.4 and figure 6.14 above, that Generic crawler search provided 130,100,80 and 160 documents for queries 'Geophysics', 'Scientist', 'Laboratory', and 'Laws' under the domain name 'Science', respectively. On the other hand, the proposed approach provided 15, 20, 10, and 20, which all are novel and filter out the redundant pages.

Result Analysis 2: As shown in table 6.4 and figure 6.14 above that Generic crawler search provided 390,360,310 and 320 documents for queries 'universe', 'nature', 'society', and 'people' under the domain name 'World', respectively. On the other hand proposed approach provide 45, 38, 20 and 28 which all are novel and filter out the redundant pages.

Result Analysis 3: As shown in table 6.4 and figure 6.14 above that Generic crawler search provided 280, 220, 180, and 170 documents for queries 'service', 'merchandise', 'manufacturing', and 'partnership' under the domain name 'Business', respectively. On the other hand, the proposed approach provides 241, 196, 150, and 145, which all are novel and filter out the redundant pages.

Result Analysis 4: As shown in table 6.4 and figure 6.14 above that Generic crawler search provided 480,430,415 and 360 documents for queries 'bus', 'car', 'truck', and 'vehicle' under the domain name 'Transport'. On the other hand, the proposed approach provides 50, 40, 30, and 42, which all are novel and filter out the redundant pages.

Memory Overhead: As shown in figure 6.15, if a page of size is 5 MB, then the generic approach memory overhead is $4385 \times 5 = 21925$ MB, and according to the proposed approach, the memory overhead is $478 \times 5 = 2390$ MB.

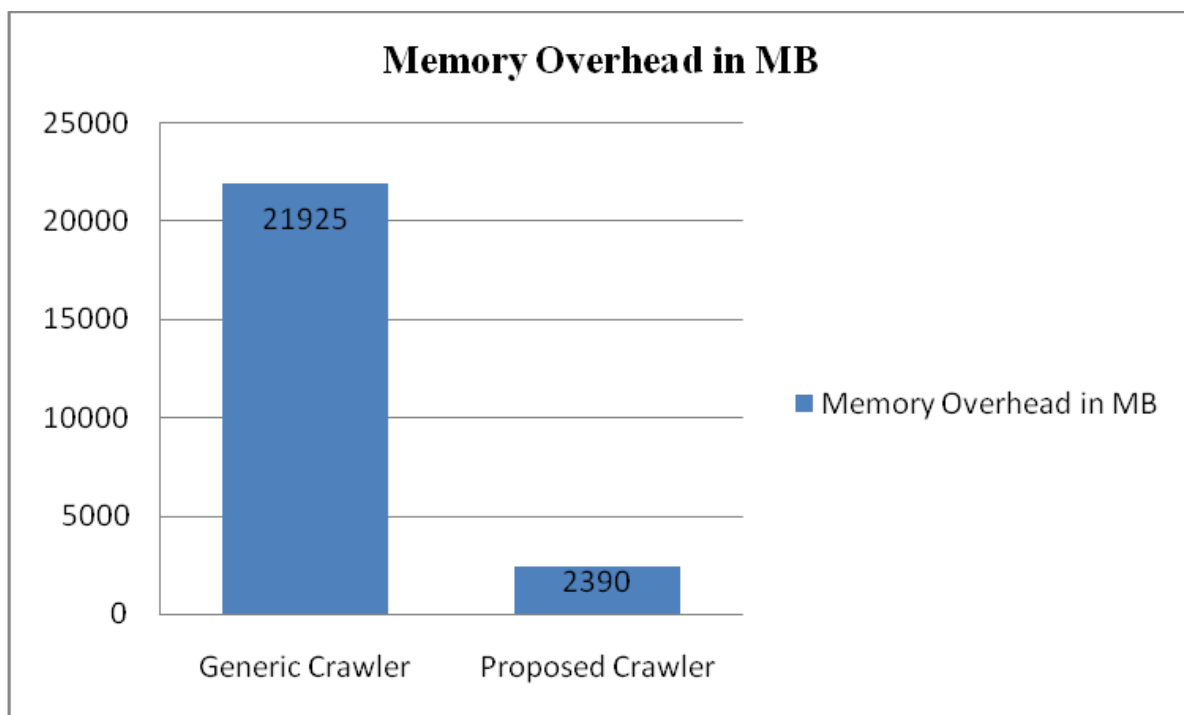


Figure 6.15: Comparison of Memory Overhead

From the above results, it has been cleared that this proposed approach provided the novel documents for a given query with minimum memory overhead and filtered out the redundant documents.

6.9 SUMMARY OF RESULTS

In this chapter a novel technique based on extractive text summarization using ontology, semantic similarity using word net 3.0, and similarity calculation using winnowing algorithm is proposed. The result after comparison with generic crawler present following inferences:

- After performing experiments with keywords/query words from different domains the proposed work gives least redundant results. The average redundancy is reduced to 88% of all the results.
- Reduced redundancy provides novel results for the prescribe search rather than replicating the previous results. This results in effective search effort.
- Memory requirement for the search results also reduce to large extent.
- One of the, main feature of this technique is that number of pages identified after the given search are very less as compared to generic technique. This results in the elimination of repeated occurrence and less memory requirement with less execution time.
- It is able to provide quality results with user need and trends.

The next chapter concludes the output of the proposed in this thesis. The future research directions are also suggested for further improving the results.

CHAPTER VII

CONCLUSTION AND FUTURE WORK

7.1 CONCLUSION

In this thesis the Design of Novelty Detection Techniques for Optimized Search Engine Results has been proposed and implemented. After the itemized investigation of existing work on Novelty Detection in text documents the specific limitations were distinguished.

The Proposed work meets the accompanying goals:

- **Redundancy free Database**

Redundancy of information both at the hour of crawling and putting away of results in the web crawler repository has been eliminated. Hashing based URL examination gives quick and simple distinguishing proof of imitated URLs.

- **Effortlessness of providing Results**

Existing calculations offer inclination to just question keywords for discovering content similarity of the website page with entered client demand. Anyway equivalent catchphrases or induced words are completely stayed away from which may influence the recovery results. The proposed work not just considers equivalent words for a given inquiry keywords yet additionally consider all gathered words that are straightforwardly and in a roundabout way identified with the keywords. Accordingly client will have the option to locate all connected data rapidly and effectively with in less time.

- **Relevant and Novel Results**

The proposed techniques utilized in this thesis join content similarity along with semantic likeness with the inclinations of domain and sort of site to be visited by the client or user. The proposed calculations are consequently proficient to produce results with client need

and inclinations. This eventually improves the general quality by giving novel results as per the client need.

7.2 ENHANCED PERFORMANCE

The result after comparison with generic crawler present following inferences:

- After performing experiments with keywords/query words from different domains the proposed work gives least redundant results. The average redundancy is reduced to 88% of all the results.
- Reduced redundancy provides novel results for the prescribe search rather than replicating the previous results. This results in effective search effort.
- Memory requirement for the search results also reduce to large extent by quickly identification of duplicate web pages.
- One of the, main feature of this technique is that number of pages identified after the given search are very less as compared to generic technique. This results in the elimination of repeated occurrence and less memory requirement with less execution time.

7.3 FUTURE WORK

This work have been discussed many issues of generic crawler based novelty detection. Anyway there are still a few issues that might be investigated in near future. The rundown of a portion of the issues is as follow

- **Working with Hidden Web**

The proposed methods are intended to work for website pages having a place with general web. These methods might be additionally stretched out to work for hidden pages also.

- **Query handling on context based**

Client setting based request might be utilized that will adapt up the need of various significance of a word in various context. It might upgrade the nature of results dependent on the client need.

- **Characteristic Language Processing ideas**

The proposed strategies should be reached out by using the automatic question suggestion techniques for accomplishing better outcomes. So Natural language preparing ideas might be utilized for proposing the client questions to give the proficient and novel recovery of web documents.

REFERENCES

- [1] Porter, Michael E., and Michael; ilustraciones Gibbs".Strategy and the Internet."(2001).
- [2] Monica Peshave and Kamyar Dezhgoshia,"How search Engines Work and a Web Crawler Application". Department of Computer Science, University of Illusion, Springfield USA.
- [3] Marios D. Dikaiakos, Athena Stassopoulou, and Loizos Papageorios, An investigation of web crawler behavior: Characterization and metrics, by Computer Communication 28(2005),880-897
- [4] Stefan Buttcher, Charles L. A Clarke, and Gordon V. Cormack. Information Retrieval: Implementing and Evaluating search Engine. MIT Press, Cambridge Mass, 2010
- [5] Saini, C.,& Arora, V.(2016, September). Information retrieval in web crawling: A survey. In Advnances in Computing , communications and Informatics (ICACCI), 2016 International Conference (pp. 2635-2643).IEEE.
- [6] Kumar, M., Bhatia, R., & Rattan, D. (2017). A Survey of web crawlers for information retrieval. Wiley Interdisciplinary Reviews: Data Mining and Knowledge discovery,7(6).
- [7] Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A. The world-wide web. Communications of the ACM. 1994 Aug 1;37(8):76-82.
- [8] E. Greengrass, "Information Retrieval: A Survey, DOD Technical Report TR-R52-008-001", (2000).
- [9] G. Salton and M. J. McGill 1983 Introduction to modern information retrieval. McGraw Hill, ISBN 0070544840.
- [10] Zhao, L., Zhang, M., & Ma, S. (2006).The nature of novelty detection, Information Retrieval, 9(5), 521-541.
- [11] Mockapetris, Paul V. "Rfc1035: Domain names-implementation and specification." (1987).
- [12] M. Franz, A. Ittycheriah, J. S. McCarley and T. Ward, "First Story Detection, Combining Similarity and Novelty Based Approach", Topic Detection and Tracking Workshop, 2001.
- [13] Shirkhoshidi, Ali Seyed, Saeed Aghabozorgi, and Teh Ying Wah. "A comparison study on similarity and dissimilarity measures in clustering continuous data." PloS one 10, no. 12 (2015): e0144059.

- [14] J. Allan, V. Lavrenko and H. Jin, "First Story Detection in TDT is Hard", Proc. CIKM, 2000.
- [15] Y. Yang, J. Zhang, J. Carbonell and C. Jin, "Topic-conditioned Novelty Detection", SIGKDD, 2002: 688-693.
- [16] J. Allan, C. Wade and A. Bolivar, "Retrieval and novelty detection at the sentence level", In proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, (**2003**), pp. 314-321
- [17] Stanford parser sentence splitting <https://www.xpdf.com/resolution/30926661.html>
- [18] N. Stokes and J. Carthy, "First Story Detection using a Composite Document Representation", Proc. HLT01, 2001.
- [19] Y. Yang, T. Pierce and J. Carbonell, "A Study on Retro-spective and On-Line event detection", Proc. SIGIR-98
- [20] T. Brants, F. Chen and A. Farahat, "A System for New Event Detection", Proc. SIGIR-03, 2003: 330-337.
- [21] T. Brants, F. Chen, and A. Farahat, "A System for New Event Detection", in Proceedings of ACM SIGIR2003.
- [22] D. Harman, "Overview of the TREC 2002 Novelty Track", TREC 2002.
- [23] I. Soboroff and D. Harman, "Overview of the TREC 2003 Novelty Track", TREC 2003.
- [24] I. Soboroff, "Overview of the TREC 2004 Novelty Track", TREC 2004.
- [25] Y. Zhang, J. Callan and T. Minka, "Novelty and Reduncancy Detection in Adaptive Filtering", Proc. SIGIR, 2002.
- [26] Tsai, F.S. D2S: document-to-sentence framework for novelty detection. Knowl. Inf. Syst.(2010)
- [27] I. Soboroff and D. Harman, "Novelty detection: the TREC experience", In proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, (**2005**), pp. 105-112.
- [28] G. Kumaran and J. Allan, "Text Classification and Named Entities for New Event Detection", in Proceedings of ACM SIGIR 2004, pp297-304.

- [29] H. Kazawa, T. Hirao, H. Isozaki and E. Maeda, "A machine learning approach for QA and Novelty Tracks: NTT system description", TREC-10, 2003.
- [30] H. Qi, J. Otterbacher, A. Winkel and D. T. Radev, "The University of Michigan at TREC2002: Question Answering and Novelty Tracks", TREC 2002.
- [31] D. Eichmann and P. Srinivasan. "Novel Results and Some Answers, The University of Iowa TREC-11 Results", TREC 2002.
- [32] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin and L. Zhao, "Expansion-Based Technologies in Finding Relevant and New Information: THU TREC2002 Novelty Track Experiments", TREC 2002. .
- [33] K.L. Kwok, P. Deng, N. Dinstl and M. Chan, "TREC2002, Novelty and Filtering Track Experiments using PRICS", TREC 2002.
- [34] M. Tsai, M. Hsu and H. Chen, "Approach of Information Retrieval with Reference Corpus to Novelty Detection", TREC 2003.
- [35] Q. Jin, J. Zhao and B. Xu, "NLPR at TREC 2003: Novelty and Robust", TREC 2003.
- [36] J. Sun, J. Yang, W. Pan, H. Zhang, B. Wang and X. Cheng, "TREC-2003 Novelty and Web Track at ICT", TREC 2003.
- [37] K.C. Litkowski, "Use of Metadata for Question Answering and Novelty Tasks", TREC 2003.
- [38] M. Zhang, C. Lin, Y. Liu, L Zhao and S. Ma, "THUIR at REEC 2003: Novelty, Robust and Web", TREC 2003.
- [39] C. Zhai, W. W. Cohen and J. Lafferty, "Beyond Independent Relevance: Method and Evaluation Metrics for Subtopic Retrieval", Proc. SIGIR-03, 2003: 10-17.
- [40] J. Allan, R Gupta and V. Khandelwal, "Temporal Summaries of News Topics," in the Proceedings of the 24th Annual International ACM SIGIR Conference, pp. 10-18, 2001
- [41] W. Dai. and R. Srihari, "Minimal Document Set Retrieval," Proc. ACM CIKM'05, pp 752-759.
- [42] R. T. Fernández and D. E. Losada, "Novelty detection using local context analysis", In proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, Amsterdam, The Netherlands, (2007), pp. 725-726.
- [43] Tang W, Tsai FS, Chen L (2010) Blended metrics for novel sentence mining. Expert Syst Appl 37(7):5172–5177

- [44] X. Li and W. B. Croft, "Improving novelty detection for general topics using sentence level information patterns", In proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, Virginia, USA, (2006), pp.238- 247.
- [45] Ng KW, Tsai FS, Chen L, Goh KC (2007) Novelty detection for text documents using named entity recognition. In: 2007 6th international conference on information, communications and signal processing, ICICS
- [46] X. Li and W. B. Croft, "Evaluating question-answering techniques in Chinese", In proceedings of the first international conference on Human language technology research, San Diego, (2001), pp.1-6.
- [47] X. Li, "Syntactic features in question answering", In proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada, (2003), pp. 383-384.
- [48] X. Li and W. B. Croft, "Sentence level information patterns for novelty detection", Ph.D. dissertation, University of Massachusetts Amherst, (2006).
- [49] Kwee, A. T., Tsai, F. S., & Tang, W. (2009, April). Sentence-level novelty detection in English and Malay. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 40-51). Springer, Berlin, Heidelberg.
- [50] Bentivogli, L., Clark, P., Dagan, I., & Giampiccolo, D. (2011, November). The Seventh PASCAL Recognizing Textual Entailment Challenge. In TAC.
- [51] Stokes, N., & Carthy, J. (2001, September). Combining semantic and syntactic document classifiers to improve first story detection. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 424-425).
- [52] Brants, T., Chen, F., & Farahat, A. (2003, July). A system for new event detection. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 330-337).
- [53] Franz, M., Ittycheriah, A., McCarley, J. S., & Ward, T. (2001). First story detection, combining similarity and novelty based approach. In 1116 2001 Topic Detection and Tracking (TDT) Workshop Report.
- [54] Larkey, L. S., Allan, J., Connell, M. E., Bolivar, A., & Wade, C. (2002). UMass at TREC 2002: Cross language and novelty tracks. *MASSACHUSETTS UNIV AMHERST CENTER FOR INTELLIGENT INFORMATION RETRIEVAL*.

- [55] Tsai, M. F., Hsu, M. H., & Chen, H. H. (2003). Approach of Information Retrieval with Reference Corpus to Novelty Detection. In TREC (pp. 474-479).
- [56] Allan, J., Wade, C., & Bolivar, A. (2003, July). Retrieval and novelty detection at the sentence level. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 314-321), ACM.
- [57] Soboroff, I., & Harman, D. (2005, October). Novelty detection: the TREC experience. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (pp. 105-112). Association for Computational Linguistics
- [58] Li, X., & Croft, W. B. (2005, October). Novelty detection based on sentence-level patterns. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 744-751). ACM.
- [59] Allan, J., Gupta, R., & Khandelwal, V. (2001, September). Temporal summaries of new topics. In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 10-18).
- [60] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." *Journal of emerging technologies in web intelligence* 2, no. 3 (2010): 258-268.
- [61] Indu, M., and K. V. Kavitha. "Review on text summarization evaluation methods." In 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS), pp. 1-4. IEEE, 2016.
- [62] Zhang, Y., Callan, J., Callan, J., & Minka, T. (2002, August). Novelty and redundancy detection in adaptive filtering. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 81-88). ACM.
- [63] Alqaraleh S., Elimination of Repeated Occurrences in Image Search Engines, Technical Report, Eastern Mediterranean University, North Cyprus, 2011.
- [64] Alqaraleh, S., & Ramadan, O. (2014). Elimination of repeated occurrences in multimedia search engines. *Int. Arab J. Inf. Technol.*, 11(2), 134-139.
- [65] Sravanthi, P., & Srinivasu, B. (2017). Semantic Similarity between Sentences. *International Research Journal of Engineering and Technology (IRJET)*, 4(1), 156-161.
- [66] Dasgupta, T., & Dey, L. (2016, March). Automatic Scoring for Innovativeness of Textual Ideas. In Workshops at the Thirtieth AAAI Conference on Artificial Intelligence.
- [67] Ghosal, T., Salam, A., Tiwari, S., Ekbal, A., & Bhattacharyya, P. (2018). TAP-DLND 1.0: A Corpus for Document Level Novelty Detection. *arXiv preprint arXiv:1802.06950*.

- [68]. Andrej Z. Broder, Steven C. Glassman , Mark S.Manasse, and Geoffrey Zweig, “Syntati Clustering of the Web”, Inproceedings of the Sixth International Conference on World Wide Web ,pp :1157-1166,1997
- [69]. Theobald , M., Siddharth ,J., and Paepcke , A. 2008. Spotsigs : robust and efficient near duplicate detection in large web collections. In SIGIR. pp.563-570
- [70]. Fetterly ,D. Manasse, M. And Najork, M. On the evolution of clusters of near duplicate web pages, In Proceedings of the first Latin American Web Congress (LAWeb), pp. 37-45, 2003.
- [71]. Hannaneh Hajishirzi, Wen-tau Yih, and Aleksander Kolcz, “Adaptive Near- Duplicate Detection Via similarity Learning”, In Proceedinnngs of the 33rd international ACM SIGIR conference on Research and development in information retrieval, SIGIR ‘ 10, pp. 416-426, 2010.
- [72] Lee, C. S., Kao, Y. F., Kuo, Y. H. & Wang, M. H. (2007). Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, 60(3), 547-566.
- [73] Henze, N., Dolog, P., & Nejdl, W. (2004). Reasoning and ontologies for personalized e-learning in the semantic web. *Journal of Educational Technology & Society*, 7(4), 82-97.
- [74] Hovy, E., & Lin, C. Y. (1999). Automated text summarization in SUMMARIST. *Advances in automatic text summarization*, 14.
- [75] Simmons, S., & Estes, Z. (2006). Using latent semantic analysis to estimate similarity. In *Proceedings of the Cognitive Science Society* (pp. 2169-2173).
- [76] Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2), 265-283.
- [77] Wibowo, A. T., Sudarmadi, K. W., & Barmawi, A. M. (2013, March). Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents. In *2013 International Conference of Information and Communication Technology (ICoICT)* (pp. 128-133). IEEE.
- [78] Gupta, S. B. (2012). The Issues and Challenges with the Web Crawlers. *International Journal of Information Technology & Systems*, 1(1), 1-10.
- [79] Meng, L., Huang, R., & Gu, J. (2013). A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1), 1-12.

- [80] Ceska, Z., Hanak, I., & Tesar, R. (2007, September). Teraman: A tool for N-gram extraction from large datasets. In 2007 IEEE International Conference on Intelligent Computer Communication and Processing (pp. 209-216). IEEE.
- [81] Hussein, A. S. (2015, May). Arabic document similarity analysis using n-grams and singular value decomposition. In 2015 IEEE 9th international conference on research challenges in information science (RCIS) (pp. 445-455). IEEE.
- [82] Ceska, Z., Hanak, I., & Tesar, R. (2007, September). Teraman: A tool for N-gram extraction from large datasets. In 2007 IEEE International Conference on Intelligent Computer Communication and Processing (pp. 209-216). IEEE.
- [83] R.C. Balabntaray, C Sharma, and M. Jha, "Document Clustering using K-means and K-medoid", Vol. 1, No. 1, June, 2013.
- [84] G. Satheelaxmi, M.R. Murty, J.V.R. Murty, and P. Reddy, "Cluster analysis on complex structured and high dimensional data objects using K-means and EM algorithm", Vol. 1, No. 1, 2012.
- [85] G. Hu, S. Zhou, J. Guan, and X. Hu, Towards effective document clustering: "A constrained K-means based approach", Information, Processing and Management, Vol. 44, No. 4, pp. 1397-1409, 2008.
- [86] C. Ding, and X. He, "K-means Clustering via Principal Component Analysis", pp. 225-232, 2004.
- [87] Likas A, Vlassis N, Verbeek JJ. The global k-means clustering algorithm. Pattern recognition. 2003 Feb 1;36(2):451-61.
- [88] Nazeer, KA Abdul, and M. P. Sebastian. "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm." In Proceedings of the world congress on engineering, vol. 1, pp. 1-3. London: Association of Engineers, 2009.
- [89] Pooja, Kumar.S, Bhatia K.K.(2019, June). Hashing and Clustering-based Novelty Detection. In SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE) – 6(6), 121-126
- [90] The anaconda website. [Online] Available: <https://www.anaconda.com/distribution/>
- [91] Jiayi, P., Cheng, C. P. J., Lau, G. T., & Law, K. H. (2008). Utilizing statistical semantic similarity techniques for ontology mapping—With applications to AEC standard models. *Tsinghua science and technology*, 13(S1), 217-222

- [92] Rus, V., Lintean, M., Banjade, R., Niraula, N., & Stefanescu, D. (2013). Seminar: The semantic similarity toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 163-168).
- [93] Cheddad, A., Condell, J., Curran, K., & McKeivitt, P. (2010). A hash-based image encryption algorithm. *Optics Communications*, 283(6), 879-893
- [94] Meadow, C. T., Boyce, B. R., & Kraft, D. H. (1992). Text information retrieval systems (Vol. 20). San Diego, CA: Academic Press.
- [95] Karkali, M., Rousseau, F., Ntoulas, A., & Vazirgiannis, M. (2013, October). Efficient online novelty detection in news streams. In *International conference on web information systems engineering* (pp. 57-71). Springer, Berlin, Heidelberg.
- [96] Alzahrani, S. M. (2009). Plagiarism Auto-detection in Arabic Scripts Using Statement-based Fingerprints Matching and Fuzzy-set Information Retrieval (Doctoral dissertation, Universiti Teknologi Malaysia).
- [97]. “A Deep Neural Solution To Document level Novelty Detection,” *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2802-2813 Santa Fe, New Mexico, USA, August 20-26-2018.
- [98]. J. Liu, W. C. Chang, Y. Wu, and Y. Yang, “Deep learning for extreme multi-label text classification,” in *SIGIR 2017 - Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017.
- [99]. H. T. Le, C. Cerisara, and A. Denis, “Do convolutional networks need to be deep for text classification?” in *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018
- [100]. X. Zhang, J. Zhao, and Y. LeCun, “Character-level convolutional networks for text classification,” in *Advances in neural information processing systems*, 2015, pp. 649–657.
- [101]. Ding, X., Li, Y., Belatreche, A. and Maguire, L.P., 2014. An experimental evaluation of novelty detection methods. *Neurocomputing*, 135, pp.313-327.
- [102]. Faria, E.R., Gonçalves, I.J., de Carvalho, A.C. and Gama, J., 2016. Novelty detection in data streams. *Artificial Intelligence Review*, 45(2), pp.235-269.
- [103]. Kerner, H.R., Wagstaff, K.L., Bue, B.D., Wellington, D.F., Jacob, S., Horton, P., Bell, J.F., Kwan, C. and Amor, H.B., 2020. Comparison of novelty detection methods for multispectral images in rover-based planetary exploration missions. *Data Mining and Knowledge Discovery*, pp.1-34.

- [104].Van Landeghem, J., Blaschko, M., Anckaert, B. and Moens, M.F., 2020. Predictive uncertainty for probabilistic novelty detection in text classification. In Proceedings ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning. ICML.
- [105].Gambhir, M. and Gupta, V., 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47(1), pp.1-66.
- [106].Dynich, Andrei, and Yanzhang Wang. "Analysis of novelty of a scientific text as a basis for assessment of efficiency of scientific activities." *Journal of Organizational Change Management* (2017).
- [107].Gupta, V. and Lehal, G.S., 2009. A survey of text mining techniques and applications. *Journal of emerging technologies in web intelligence*, 1(1), pp.60-76.
- [108].Karami, A., 2019. Application of fuzzy clustering for text data dimensionality reduction. *International Journal of Knowledge Engineering and Data Mining*, 6(3), pp.289-306.
- [109]. Meng Y, Zhang Y, Huang J, Xiong C, Ji H, Zhang C, Han J. Text classification using label names only: A language model self-training approach. *arXiv preprint arXiv:2010.07245*. 2020 Oct 14.
- [110]. Fernandes P, Allamanis M, Brockschmidt M. Structured neural summarization. *arXiv preprint arXiv:1811.01824*. 2018 Nov 5.
- [111] Sushil Kumar, Komal Kumar Bhatia, Ishuka, "A Novel Approach for Novelty Detection using Extractive Text Summarization", in *Journal of Emerging Technologies and Innovative Research (JETIR)*, Volume 6, Issue 6 , ISSN(online): 2349-5162, June 2019, pp. 141-154
- [112] Sushil Kumar, Komal Kumar Bhatia, "Semantic Similarity and Text Summarization based Novelty Detection", in *SN Applied Sciences*, Volume 2, Number 2, Feb 2020.

BRIEF PROFILE OF THE RESEARCH SCHOLAR

Sushil Kumar did his B.Tech (Computer Engineering) from N.I.T Kurukshetra (Deemed University), Kurukshetra in 2002 and M.E (Computer Science & Engineering) from Punjab University Chandigarh in 2007. Currently he is working as Assistant Professor in the department of Computer Engineering at J.C Bose University of Science & Technology, YMCA, Faridabad India. He is having a total of 16 years of teaching experience. His area of research includes Internet technologies, Information Retrieval System and Computer Organization & Architecture. He has authored papers in various national and international journals.

LIST OF PUBLICATIONS

Publication in International Journal

1. Sushil Kumar, Komal Kumar Bhatia ,“Clustering Based Approach for Novelty Detection in Text Documents”, in Asian Journal of Computer Science and Technology(AJCST), Vol.8 No.2, ISSN(online): 2249-0701,June 2019, pp. 121-126.
UGC Approved.
2. Sushil Kumar, Komal Kumar Bhatia, Ishuka, “A Novel Approach for Novelty Detection using Extractive Text Summarization”, in Journal of Emerging Technologies and Innovative Research (JETIR), Volume 6, Issue 6 , ISSN(online) : 2349-5162, June 2019, pp. 141-154
3. Sushil Kumar, Komal Kumar Bhatia, Pooja, “Hashing and Clustering based Novelty Detection”, in SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE), Volume 6 Issue 6 , ISSN(online): 2348 – 8387, June 2019, pp.1-9
4. Sushil Kumar, Komal Kumar Bhatia,“ Semantic Similarity and Text Summarization based Novelty Detection”, in SN Applied Sciences, Volume 2, Number 2, Feb 2020 indexes in **Emerging Sources of Citation Index (ESCI), published by Springer.**

Publication in International Conference

1. Sushil Kumar, Komal Kumar Bhatia, “Document to Sentence level Novelty Detection”, 50th Golden jubilee international annual convention of Computer Society of India (CSI-2015) theme Digital Life, organised by BVICAM New Delhi, 2nd to 5th December 2015. **Conference proceedings published by Springer.**

Publication in National Conference

1. Sushil Kumar, Komal Kumar Bhatia, Ashutosh Dixit, “Search Engine Tools: A Review”, in National Seminar of Advancement in Technology, July 2012 at IET Baddal, Punjab.
2. Sushil Kumar, Komal Kumar Bhatia, “Extractive Text Summarization Using Regression Model”, in National Conference on Role of Science and Technology towards: Make in India, March 05-07, 2016 at YMCAUST, Faridabad .
3. Sushil Kumar, Komal Kumar Bhatia, “ A Novel Approach for Novelty Detection via Topic Modelling”, in National Conference on Advanced in Mathematics & Computing, 1st to 2nd May 2017 at YMCAUST, Faridabad.