Supervised Text Classification for News Filtering and Summarization on the Web

THESIS

SUBMITTED TO

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

LUCKNOW

BABASAHEB
BHIMRAO
AMBEDKAR
UNIVERSITY

LUCKNOW

प्रज्ञा शील करुणा

ESTABLISHED 1996

FOR THE DEGREE OF

Doctor of Philosophy

IN

COMPUTER SCIENCE

Submitted by

Chandrakala Arya

Enrolment No. 957/13

Supervisor

Prof. Sanjay Kumar Dwivedi

DEPARTMENT OF COMPUTER SCIENCE

SCHOOL FOR INFORMATION SCIENCE & TECHNOLOGY

BABASAHEB BHIMRAO AMBEDKAR UNIVERSITY

(A CENTRAL UNIVERSITY; NAAC- 'A' GRADE)

VIDYA VIHAR, RAEBARELI ROAD, LUCKNOW-226 025 (U.P.), INDIA

2018

CANDIDATE'S DECLARATION

I hereby declare that I have completed research work for the full time prescribed and that the thesis embodies the results of my investigation conducted during the period I worked as Ph.D. research scholar. I further declare that to the best of my knowledge the thesis does not contain part of any work submitted for the award of any degree either in this Institute/University or any other Institute/ University.

(Chandrakala Arya)

Research Scholar

CERTIFICATE

This is to certify that the thesis titled "**Supervised Text Classification for News Filtering and Summarization on the Web** " submitted by **Ms. Chandrakala Arya** is an original research work and has not been previously submitted in part or full for the award of any other degree or diploma to this or any other University.

The thesis submitted to Babasaheb Bhimrao Ambedkar University, Lucknow satisfies all the requirements as stipulated in the *Doctor of Philosophy (Ph.D.) regulations-1999 as amended in 2010* and it is fit for submission and evaluation for the award of the degree of Doctor of Philosophy of the University.

Date: **Supervisor**

Head of the Department

ABSTRACT

In recent times, text summarization has gained enormous attention from the research community. Text Summarization is one of the applications of natural language processing and is becoming very important in the field of information retrieval. Especially, during the last two decades, a number of efforts have been made by researches to generate good informative summary. The task of text summarization can be defined as the process of automatically creating a summarized form of given documents by retaining its important information. It also helps users to grasp main concepts of information sources in a short time.

Meanwhile, the present trend in text summarization is focusing more on the domain of News Summarization. Early work in summarization started with single document summarization. Single document summarization produces summary of one document. As research progressed, and due to large amount of information on web, multi document summarization emerged. Multi document summarization produces summaries from many source documents on the same topic or same event. News Summarization system however lacks the capability to deal with multi-document news summarization due to the content redundancy. Therefore, it is this perspective with which we initiate our study in News Summarization by analyzing and presenting different approaches used in the research area.

With this background, we propose a news summarization system based on the main phases as news web page classification, filtering using content extraction, keyphrase extraction and finally summary is generated by sentences ranking and redundancy reduction.

In news web page classification, news web pages have been classified from the non-news web pages by extracting the three attributes content, structural and URL and used Naïve Bayes classifier for the classification. Naïve Bayes classifier is also compared with the SMO and J48 classifiers for the same dataset, and results show that it gives good results than the remaining two. After that important contents have been extracted from the correctly classified news web pages. It depends on the tokenization of the HTML page and these tokens construct the tag tree, which is used to find matching patterns and filter out shared token sequences until the relevant content is extracted. Then, extracted relevant content is used for the keyphrase extraction from the news articles. Keyphrases can be a single word or a combination of more than one word that

represents the news article's important concept. Our proposed approach of keyphrase extraction is based on the identification of candidate phrase from the news articles, and chooses the highest weight candidate phrase using the weight formula. Weight formula includes features such as TFIDF, phrase position and construction of lexical chain to represent the semantic relations between words using WordNet. Proposed approach shows good results compared to the other existing approaches.

We have focused on sentence ranking and reducing similarity from the previous phase to the next one. Extracted keyphrases are used as a feature direct keyphrase match and has been combined with the other features as, matching terms, sentence position, and sentence length to calculate sentence weight for the sentence ranking. For the similarity reduction, cosine similarity measure is used. The experiments results showed satisfactory ROUGE values and provides accurate and precise summary of the multi-documents news articles.

The work carried in this thesis clearly indicates that the performance of News Summarization system can be improved significantly with the correct classification of news web pages, correct content extraction, keyphrase extraction and redundancy reduction.

ACKNOWLEDGMENT

This thesis becomes a reality with the kind support and help of many individual. I would like to extend sincere thanks to all of them.

First and Foremost I would like to Thank **God.** You have given me the power to believe in myself and pursue my dreams. I could never have done this without the faith I have in you, the Almighty.

The award of degree Doctor of Philosophy is one of the hardest deserving achievements. People struggle for it and achievement not easily found. During the entire research works some valuable people conceived their enormous positions in my heart. In this regard, I am grateful to the University and express my deep sense of gratitude to its **Hon'ble Vice-Chancellor** for delivering this great opportunity to me.

I would like to extend sincerest gratitude with heartfelt thanks towards my Supervisor, **Prof. Sanjay K. Dwivedi, Department of Computer Science, Babasaheb Bhimrao Ambedkar University, Lucknow**, for his support. He encouraged me to explore on my own while constantly providing insightful advice and prepared me for academic life, he taught me about research and always made me believe that I could succeed, he patiently corrected my papers even when he had millions of other things waiting. Working with him, I learned the value of a vision, and persistence in occurring it. I am most grateful to him.

I convey my sincere thanks to **Head** and all other faculty members of department for their motivation and support during the research. I would also like to show my gratitude to the Department of Computer Science, for the providing the healthy and pleasant environment required for good quality research.

I would like to thank all administrative and supporting staffs of the University for providing the comfortable environment and help.

Specially thank should be given to financial supporting body, University Grants Commission that provided me fellowship for this research work.

I am heartily grateful to my senior research group members Dr. Parul Rastogi, Dr. Ajay Kumar Bharti, Dr. Pramoad P. Shukdeve, Dr. J.N.Singh, Dr. Rajesh Gautam, Dr. Anand Kumar & Dr Vaishali Singh for great help and time to constant encouragement throughout the research.

I particularly would like to deeply appreciate the great help of research scholars Ganesh Chandra, Bhupesh Rawat, Shweta Vikram, and Sachin Sahu, who always kept the healthy research environment and extended their full cooperation during my research. They have given me support, and joyful university life.

I want to thank my family because; I could not have done this without their love and support. My father Mr. Rajendra Ram Arya and my mother Kanti Devi, my sisters Rashmi, Shobha and Kanchan for their moral and emotional support.

And finally, I thanks to my brother Mr. Naveen Krishn Nagar for showing love and support in both times of great happiness and in times of hardness.

I cannot list the names of all people who I indebted to but thanks to all valuable persons, who have given me enormous support and inspiration directly or indirectly during my research work.

Chandrakala Arya

Date:

Place: Lucknow

LIST OF FIGURES

LIST OF TABLES

CONTENTS

Chapter 1

Introduction

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION

This chapter introduces general concept of automatic text summarization, General architecture of automatic text summarization, Taxonomy and approaches of automatic text summarization, domain specific summarization, and need of news summarization, news filtering, Text classification, Text classification approaches, Supervised Text Classification, and major challenges in summarization. It also provides a formal description about the objectives considered for this study. The chapter concludes by rational of research and organization of thesis.

1.2 AUTOMATIC TEXT SUMMARIZATION

As the internet and online information services are growing day by day, there is a huge amount of information is available that can cause information overload problem. Therefore automatic text summarization is required. It is the process of filtering the most important information from a source or many different sources to reduce the amount of information in a textual document while preserving the most important information and produce a concise summary.

Automatic Text summarization is known as a vital research field by numerous organizations such as in DARPA [1] (United States), the European Community and Pacific Rim. It is also been increasingly used in the commercial sector such as in BT's ProSum [2] (telecommunication industry), in Oracle's Context (data mining of text databases), and filters for web based information retrieval.

Summaries of documents have been used for presenting the most relevant information from one or several documents for a long time. This is a task traditionally performed by humans, since it requires understanding of natural language as well as an understanding of what

information the readers of the summary are interested in. Task performed by humans are however expensive, and any text which is to be summarized must first be read by a human. This is not a big problem for single document summarization, since it is likely that the document has been written by a human; the amount of extra work and time spent on summarizing the document is small compared to the amount of work and time spent on writing it.

Good summaries of documents should contain the most relevant information from the documents which were summarized. There have been several attempts at creating good summarizers of multiple documents, using several different methods and assumptions. This report presents a search based approach to the problem of summarization of multiple documents. The concept of automatic summarization began in the late 1950s by Luhn [3]. Luhn's method uses term frequencies to select important sentences for the summary. Luhn used the idea based on the knowledge that in the document, significant words those carries most information are neither frequent nor rare. Therefore it is important to use the frequency of significant words and their distance in the sentence for sentence ranking and choose highly ranked sentences as a result.

Ten years later Edmundson's [4] shows remarkable progress by introducing hypothesis that concern the features like high information value of title phrases, sentence position and sentence containing cue words and phrases.

Further in 90s, Jones [5] defines summary as a content reduction of source document by selecting and generalizing the important information from the source. It is a summarize version of document including only the important information.

1.3 GENERAL ARCHITECTURE OF AUTOMATIC TEXT SUMMARIZATION

Based on the discussion of the automatic text summarization, it is evident that there are some common key activities, which formulate a typical automatic text summarization system. An overview of such activities is provided in Figure 1.1. These activities are usually executed in

a sequence. However, depending upon the technique being followed, one or more activities may be added or removed.

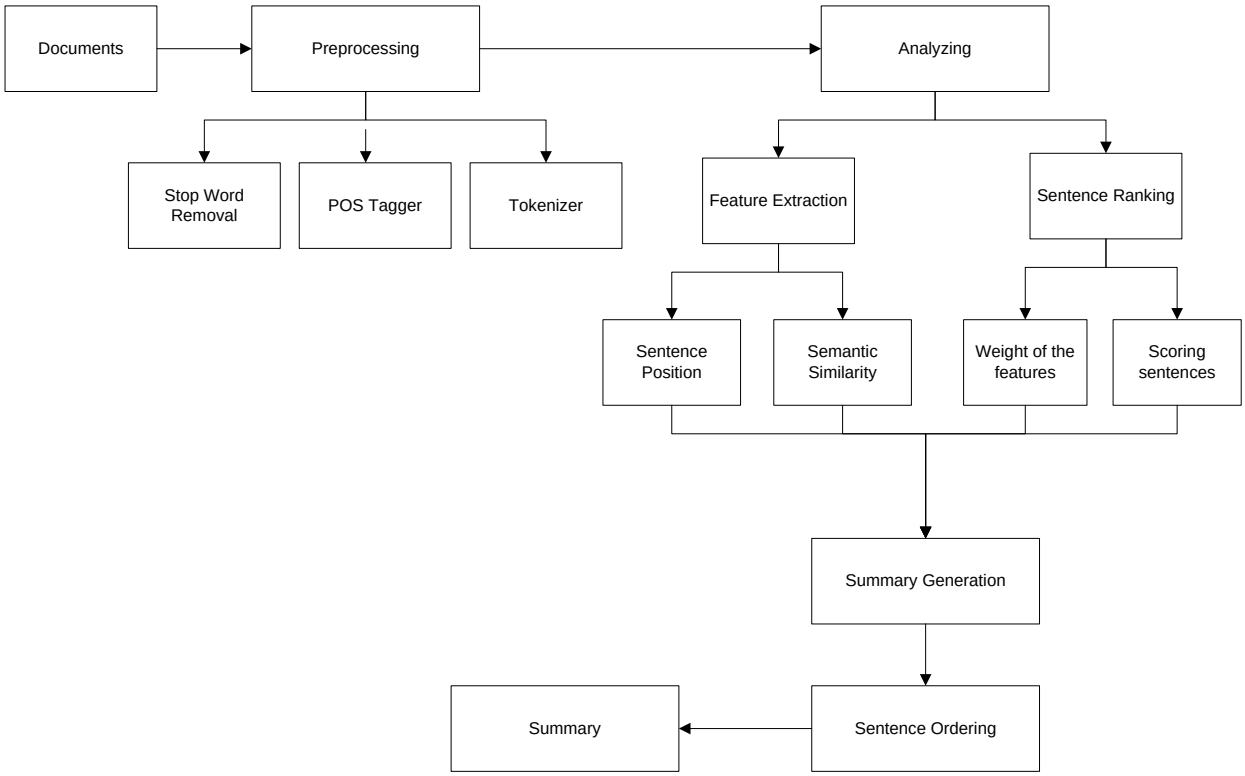


Figure1.1. General Architecture of Automatic Text Summarization [6]

The general architecture of automatic text summarization is divided into three steps as preprocessing, analyzing and summary generation.

1.3.1 PREPROCESSING

Pre-processing is a primary step to load the documents into the system. It cleans the source documents, annotating the sentences and extracts the features. This includes tokenizer, stop word removal and POS tagger. In this step, tokenize the whole article into sentence string and remove unnecessary elements such as HTML tags, news advertisements, and table numbers from the document for relevant content extraction. In further processing POS (Part–of–speech) tagged each sentence and stop words are removed from the document.

1.3.2 ANALYZING

After documents preprocessing, sentences are scored based on the extracted features. In this step, the extraction and analyzing of features are done. Basically the features such as sentence position, semantic similarity, weight of the features and sentence scoring is done. Basically the sentence position in a document can play an important role in finding the sentences that are most related to the topic of the document [7]. Semantic similarity is calculated by finding an average of the score for each noun and verb in both of the sentences. The score of each sentence is computed by the linear combination of the weighted features.

1.3.3 SUMMARY GENERATION

In this step, summary is generated by choosing the highest weight most important sentences in the document and orders the sentences in chronological order to insure the readability of the generated summary. Multi-document summaries are also generated in a similar manner by calculating sentence score of each document separately and then choosing the highest scoring sentences from all documents to generate summaries.

1.4 TAXONOMY OF AUTOMATIC TEXT SUMMARIZATION

Taxonomy of summarization depends on what is summary intend for, how is the input, output etc. Jones [8] in 1999, proposed one of the most well-known existing taxonomies, where three classes of context factors that influence summaries are taken into consideration as input, purpose and language factors shown in Table 1.1. Input factors deal with aspects related to the source, such as genre, language, or register. The second ones, purpose factors, include audience and use, for example literary reviews or emergency alerts. Finally, language factors, focus on the style and coverage, and are normally driven by purpose factors.

Figure 1.2 shows the taxonomy of automatic text summarization based on the different factors.

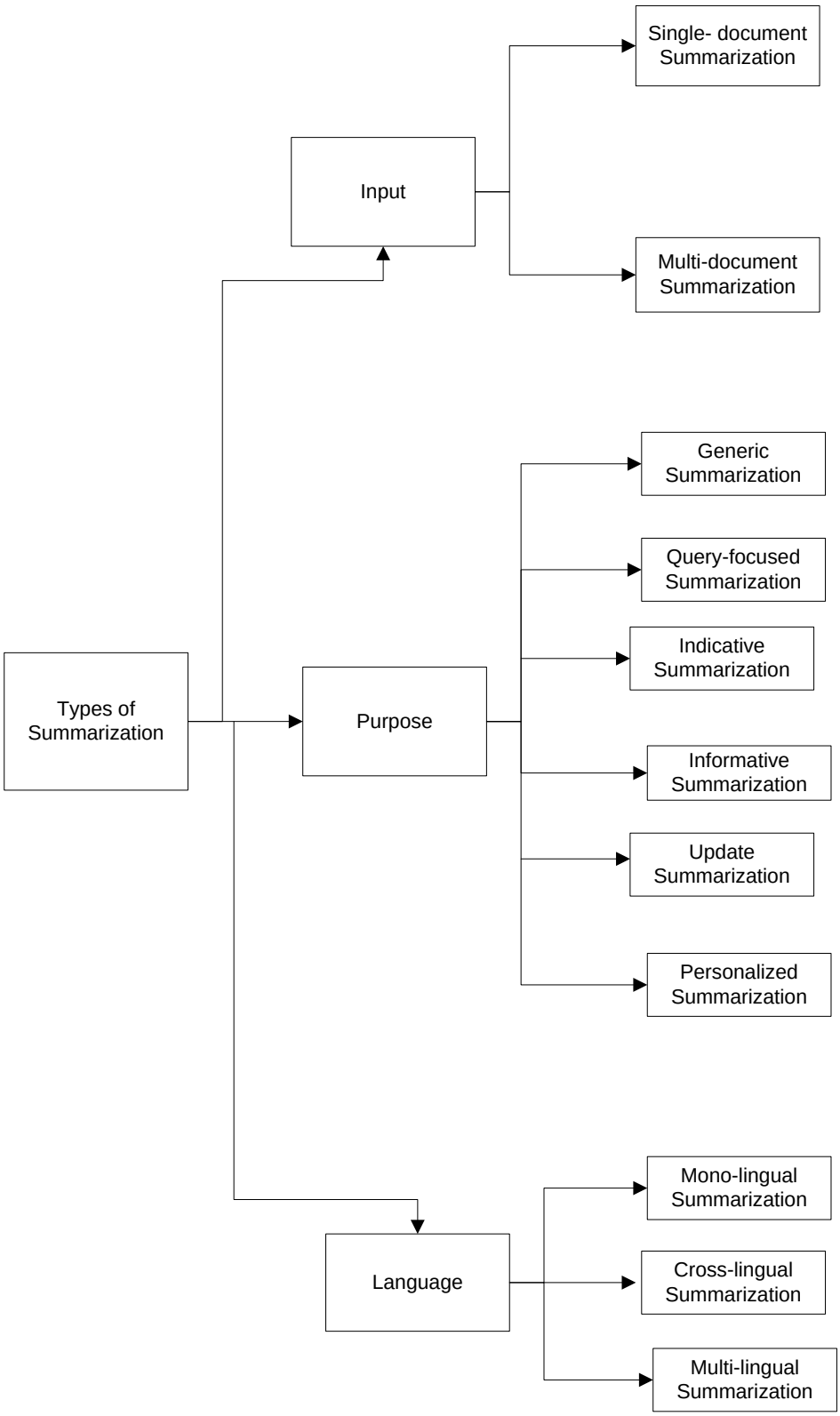


Figure 1.2. Taxonomy of Automatic Text Summarization

Table 1.1: Taxonomy of Automatic Text Summarization

Factors	Summarization Type
Input	Single-document, Multi-document
Purpose	Generic, Query-focused, indicative, informative, update, personalized
Language	Mono-lingual, Multi-lingual, cross-lingual

1.4.1 NUMBER OF INPUT DOCUMENTS BASED SUMMARIZATION

Summarization can be classified based on the number of input documents a system accepts as input [9]. Summary is generated from a single document in single-document summarization whereas in multi-document summarization, many documents are used for generating a summary.

1.4.1.1 Single-Document Summarization

In single-document summarization, the summarization is generated from only one document; the sentences can be selected according to its weight. A single-document summarization system produces a summary from only one document. Initially, despite the research on alternatives to extraction, today most of the work still depends on sentences extraction from the source document to generate a summary. To generate a single output that summarizes the salient points across multiple documents is more difficult. Since the documents are related by a common topic, they likely contain similar content; thus a system cannot simply concatenate many single document summaries together.

1.4.1.2 Multi-Document Summarization

Multi-documents summarization is an important field of the Natural Language Processing (NLP) [10]. It would be of great use given the enormous amount of news published daily online. It produces a single summary of a set of related source documents. Different from single document summarization, sentences of multi-document summarization are selected from different documents. It is important for multi-document summarization to determine a strategy to order the sentences.

The summary is generated by choosing the most important sentences in the document (ones with the highest score) and arranging them in chronological order. Multi-document summaries could be generated similar way by calculating scores of sentences in each document separately and then choosing the highest scoring sentences from all documents. Different from single document summarization, sentences of multi-document summarization are selected from different documents. It is important for multi-document summarization to determine a strategy to order the sentences.

1.4.1.3 Difference between Single-Document and Multi-Document Summarization

By the number of input documents, automatic summarization is divided into single-document summarization and multi-document summarization. There are different complications regarding both [11], including: (1) Levels of redundancy which is significantly higher in multi-document summarization, (2) Compression ratio, which is usually larger on multi-document summarization because of redundant contents being omitted, and (3) Speed of summarization is high for single-document summarization than Multi-document summarization. (4) Co-reference, presents grater challenges for multi-document than single document summarization. The comparison of single-document and multi-document summarization has been shown in Table 1.2.

Table 1.2: Comparison between Single-Document and Multi-Document Summarization

Parameters	Single-Document Summarization	Multi-Document Summarization
Levels of redundancy	Lower	Higher
Compression ratio	Smaller	Larger
Speed	High	Low
Sentence Ordering	Trivial	Acute
Co-reference	Trivial	Acute

1.4.2 PURPOSE BASED SUMMARIZATION

Summarization can be classified based on the purpose as generic summaries and query-focused summarization (also known as user-focused or topic-focused), Indicative and informative summarization, update and personalized summarization.

1.4.2.1 Generic Summarization

A broad community is addressed by the generic summarization [12]. There is no focus on special needs because the summarizer is not targeting any particular group. A Generic summary should cover the major topics of the article as much as possible and also keep the minimum redundancy. There are two methods are present for creating generic summaries, the first one is sentence selection based on relevance measure and the second is latent semantic analysis.

a) Relevance Measure based Summarization

For relevance measure firstly the article is decomposed into individual sentences and these sentences are used to form the candidate sentence set S and create the weighted term frequency vector 'V_i' for each sentence i Є S, and the weighted term frequency vector 'A' for the whole article [13]. After that for each sentence i Є S, relevance score between 'V_i' and 'A' is computed then select the sentence 'N' that has the highest relevance score and add it to the summary. Delete N from S and eliminate all the terms contained in 'N' form the article and weighted term frequency vector 'A' for the article is recomputed and generate the final summary.

b) Latent Semantic Analysis based Summarization

According to Gong and Liu in 2002 [13] Latent Semantic Analysis (LSA) is based on the idea of latent semantic indexing and applied the Singular Value Decomposition (SVD) for generic summarization. The process of LSA start with the construction of a terms by sentence matrix B = {B_1, B_2,……B_n} with each column vector B_n representing the weighted term frequency vector of sentence n in the document under consideration. If there are a total of p terms and n sentences in the document, then create a p × n matrix B for the document where without loss of generality p>=n; the SVD of B is defined as [46]:

$$B = U \sum V^T \tag{1.1}$$

Where U = [u_{ij}] is an p × n column orthogonal matrix whose columns are called left singular vectors, Σ = diag (σ_1, σ_2, ……., σ_n) is an n × n diagonal matrix whose diagonal elements are non-negative singular values sorted in descending order, and V = [v_{ij}] is an n × n orthonormal matrix whose columns are called right singular vectors.

1.4.2.2 Query-Focused Summarization

In query focused summarization, the aim is to summarize only the information in the input document(s) that is relevant to a specific user query [8]. To generate a useful summary in this context, an automatic summarizer needs to take the query into account as well as the document. The summarizer tries to find information within the document that is relevant to the query or in some cases, may indicate how much information in the document relates to the query. Query-focused summaries are query biased. These summaries are used for answering such questions as whether a given article is relevant to the user's query and if the article is relevant, which parts or sentences of the article is more relevant than others, relevant sentences are scored based on the frequency of query words and phrases. Those sentences are scored high, which contain the query phrase rather than the single query words. There is also a limitation in query-focused summaries, they do not provide an overall sense of the article content and hence are not appropriate for content overview.

1.4.2.3 Indicative Summarization

Indicative summarization is utilized to demonstrate what topics are addressed in the source document. As a result, they can give a brief idea about the original content, without carrying precise content. It gives the condensed information on the main topics of a document and should save its most imperative parts that are regularly used as the end some portion of IR system, being returned by search system instead of full document. The main aim of the indicative summaries is to aid a user with deciding whether the original document is worth reading and the length of the indicative summaries should be range between 5 till 10% of the complete text [14].

1.4.2.4 Informative Summarization

Informative summarization covers all significant information of the document. It acts as a substitute for the source document and provides some shortened version of the content, mainly by gathering relevant information in a concise form while reducing information volume. The length of it is typically 20-30% of the original document [14].

1.4.2.5 Update Summarization

For the long-running events with a dynamic flow of information typically requires concise on-topic update in a timely manner. Therefore summarization is also observed as time sensitive. Update summarization takes this time-dimension into account, and produces an incremental summary which contains the most salient and evolving information from a collection of input documents, leaving from the assumption that the user has prior knowledge about the event and has read previous documents on the topic. The summary is expected to convey the most important developments of an event beyond what the user has already seen, i.e. only new information not covered inside an initial update summary.

1.4.2.6 Personalized Summarization

The aim of personalization is to adapt the content presented to an individual user or the way they access the content based on their potential interests. Readers are required to select their possible interests before summary generation so that the chosen topic has priority during summarization.

Personalized summarization is known as a process of summarization that preserves the specific information that is relevant for a given user profile, rather than information that truly summarize the content of the news articles. The potential of summary personalization is high, because a summary that would be useless to decide the relevance of a document if summarized in a generic manner, may be useful if the right sentences are selected that match the user interest. To build a personalized or user-adapted summary a representation of the interests of the corresponding user is required. This representation may vary from a set of keywords to a complex user profile where the information needs of the user are represented according to several systems of reference [15] [16].

Personalization summarization approaches are mainly categorizes into four categories. These are:

- Content based
- Collaborative
- Demographic

- Hybrid

a) Content Based

Content based approach depends mainly on the contents of the items in a collection. It relies on the user's profile analysis model [17]. It relies on the profile built up by evaluating the content of individual user rated items given by users and uses the content of articles and rating to create a profile to contrast with other non- rated items which are based on the contents similar to user queries to create personalized summaries.

b) Collaborative Approach

It is a method of making automatic predictions about user interests by using a database of preferences for items or articles another user may like [18]. Those users who prefer the same thing in the past are prone to have similar preferences later on. Enhancement of the quality of collaborative approach depends on the user's ratings; it ignores the data that can be separate from semantic content.

c) Demographic Approach

The goal of the demographic approach is to categorize the user by demographic information like age, gender and education for the identification of users, In other words, demographic information concern with users who have rated particular article highly. The Demographic approach may not require the history of user rating like collaborative and content-based techniques; it is the advantage of this approach [19].

d) Hybrid Approach

Hybrid approach endeavors to consolidate different techniques to remove their disadvantages commonly. Most hybrid methods applied user profile and description of items to find users who have related interests and after that used collaborative filtering to make predictions. In such systems, a careful selection of features is required [19].

1.4.3 LANGUAGE BASED SUMMARIZATION

Mono lingual system only accepts documents with specific language and output is based on that language only. Multi-lingual systems can accept documents in different languages and produce summary of different languages. Cross-lingual system works with multi-linguality.

1.4.3.1 Monolingual Summarization

In monolingual Summarization, input and output language is same for example if a summarization system produces an English summary from one or more documents in English are lead to a case of monolingual summarization. FarsiSum [20] is a type of mono lingual text summarization systems.

1.4.3.2 Cross-Lingual Summarization

If the summary is produced in English but the original documents are in Hindi, the summarizer would deal with cross-linuality, since the input and output languages are different. Cross- lingual summarization was studied on the Johns Hopkins research workshop in the 2001, where evaluation resources and summarization algorithms for English and Chinese were developed [21].

1.4.3.3 Multi-Lingual Summarization

Towards the growing trend of multiliguality on the Internet require text summarization techniques that work equally well in multiple languages. Automated summarization methods can be defined as language-independent, if they are not based on any language-specific knowledge. Such methods can be used for multi-lingual summarization defined by Mani [22] in 2001 as processing of several languages with summary in the same language as input. Multi-lingual summarization is useful when the documents are in several different languages on the same event and published on a particular day to generate a summary. Multilingual summarization introduces the problem of translating the documents to the language of the summary.

Multi-lingual Summarization Evaluation in 2005, focused on summarization from mixed input in Arabic and English, where the challenge was to generate output from automatic

translations [23]. There is various research projects exists for the multilingual summarization like the SUMMARIST [24] research project has produced a summarizer available for Korean and Spanish, MUSE system [25] for English, Arabic, and Hebrew. Likewise, various researchers claim that their summarizers to be language independent [26] [27 [28].

1.4.4 COMPARATIVE ANALYSIS OF PURPOSE AND LANGUAGE BASED SUMMARIZATION

Different types of summarization that have been so far discussed in the previous section have their own capabilities. The comparison of purpose and language based summarization types have been shown in Table 1.3.

Table 1.3: Comparison between Purpose and Language Based Summarization

Summarization Type	Summary generation	Selection of summary sentences from the original article	Method used	Application areas	Systems	Level of analyzing the article
Generic	Summary is generated by covering the major topics of the article as much as possible and also keep the minimum redundancy.	Based on relevance measure and latent semantic measure	Relevance measure and latent semantic measure	Any areas	SUMMARIST	Topics of article
Query focused	Summary is generated by a user need or a query.	Based on the frequency count of the query phrases in the sentence.	Statistical methods	Any areas	WebSUMM	Terms of article
Indicative	Summary is composed by the abbreviated information on the main topics of a	Based on matching specific patterns	Information retrieval method	Any areas	SUMUM	Topics of article

	document.					
Informative	The summary is generated by the shorten version of the content while retaining important details.	Based on informative marker and satisfies an informative pattern.	Information retrieval method	Any areas	Cut-and-paste	Description of entity
Update	The summary is generated by time – dimension and produce incremental summary.	Based on TF-ISF and the number of keywords	Statistical methods	Real time application areas	Pythy	Article topic
Personalized	The summary is generated based on the specific information that is relevant for a given user profile	Based on the user interest	Content, Collaborative, Demographic, Hybrid	Recommendation application areas	NewsDude	User's interest
Mono-lingual	The system generates the summary in the same language as the input language.	Based on the sentence position, similarity to the title	SVM, NetSum	Any area	FarsiSum	Topics of article
Cross-lingual	The system can accept a specific language source text and generate summary in another language.	Based on position and cue phrase	Statistical methods	Domain-specific texts	MUSI	Deeply analyze extracted sentences
Multi-lingual	Summary is generated from different document that	Based on cue phrases, matching user query and	MUSI approach	Any area	Columbia Newsblaster	Terms of article

	are in several different languages on the same event and published on the same day.	sentence position within the text.				

1.5 APPROACHES TO AUTOMATIC TEXT SUMMARIZATION

The aim of text summarization is to extract content from an information source and present the most important content to the user in a condensed form and in a manner sensitive to the user's or an application's need [29]. There are two approaches of summarization depending on how they are done [11]. *Extractive* summarization produces a summary by extracting a subset of the sentences related to the main topic from source documents, then concatenating them to produce the final summary [30]. On the other hand, an *abstractive* summary [31] is written to express the main information by modifying phrases, sentences from the source document that do not appear in the original document, usually, abstractive summarization requires heavy machinery for language generation and is difficult to do properly, because one has to understand the point of a text which requires semantic analysis, inferential interpretation. Therefore, most of the summarization researches today focus on extractive summarization. We, here present taxonomy for categorizing various approaches followed by automatic text summarization so far as shown in Figure 1.3.

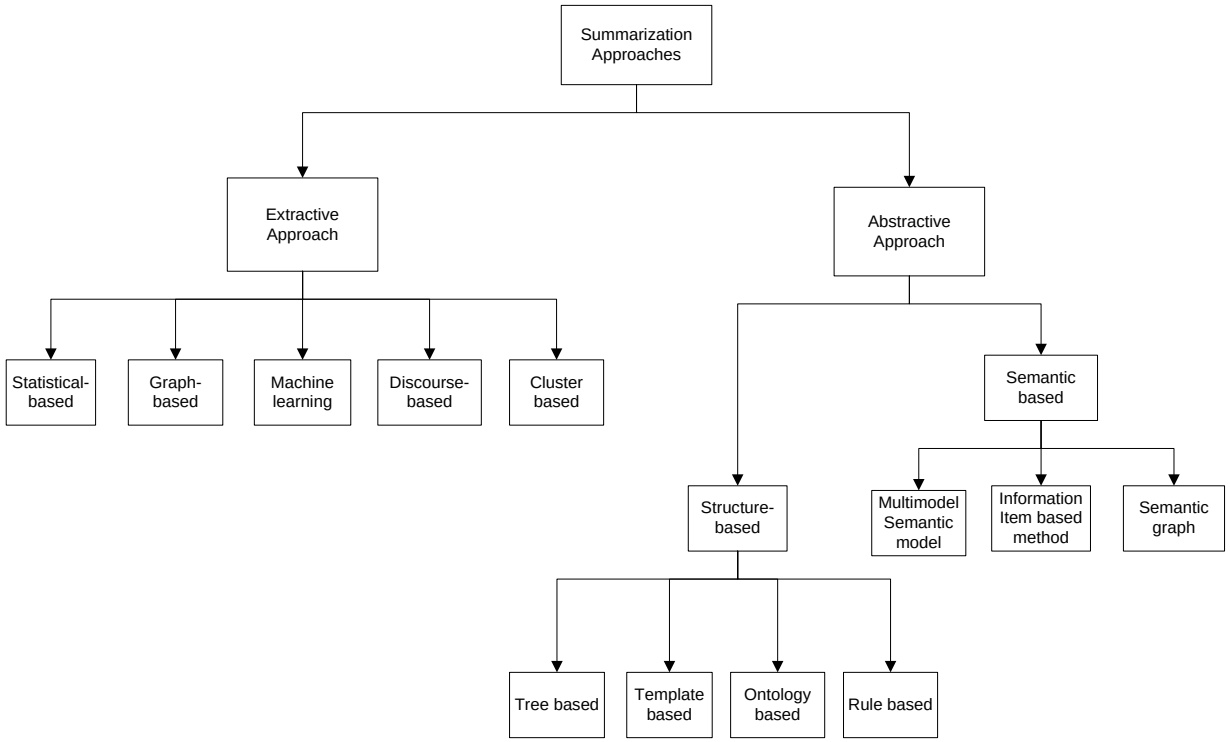


Figure 1.3. Automatic Text Summarization Approaches

1.5.1 EXTRACTIVE SUMMARIZATION APPROACH

The extractive summary consist to make procedure by identifying important section of the paragraphs; to produce a summary from the source document without changing it, extract and collect the salient sentences or phrases from the source text [32]. Extraction methods emphasis on to determine the salient sentences by matching phrasal patterns or by considering the source text lexical and statistical relevance. Original sentences from the source document are concatenated in synthesis. Linear weighting model is adopted by the most of the methods. In the analysis phase of this model, weight of each sentences are calculated according to the features like sentence location in the source document, frequency in the source document, cue phrases appearance, and statistical significance matrix. In general, the strategic flow of automatic text summarization based on extractive approach as shown in Figure 1.4 can be concise as:

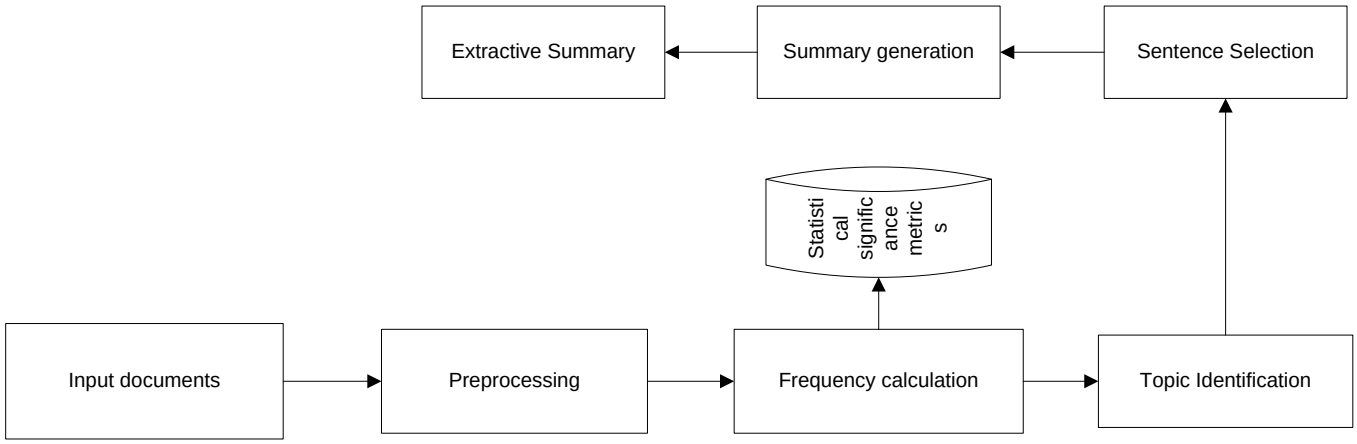


Figure 1.4. Workflow for Extractive Summarization Approach

Most of research in automated summarization use extraction approach to produce summarization and the first step in this method consist to identify important features such as sentences length and location [33].Other technique such as the number of numerical data [34], the number of proper nouns [35], the number of words occurring in title [36] and the term frequency [37] are also used.

In our work we used extractive summarization systems to produce short, paragraph length summaries and describe the performance of summarization systems. These summarizers identify the most important sentences from the source document as input, which can be either a single document or a group of related documents, and concatenate them together to structure a summary. Extractive summarization can be done using variable methods such as:

1.5.1.1 Statistical based Method

In statistical-based approach sentence has been selected based on the frequency of word, indicator phrases and other features irrespective of the words sense [38]. The idea behind these methods is depends on the most obvious indication of the document content. There are numerous methods are present for determining the main sentences such as: The Title Method [39], The Location Method [40], The Aggregation Similarity Method [41], The Frequency Method [37], and TF- Based Query Method [39].

1.5.1.2 Graph-based Method

Graph based representation [42] of documents provides a subject or title identification method that can be obtained by the common preprocessing steps such as stop word removal and stemming, document's sentences are represented as nodes in an undirected graph. If the two sentences share the common words then they are connected with an edge, on the other hand their similarity is above some threshold. In the document, high cardinality nodes are the important sentences and carry higher preference to be included in the summary. For query-specific summaries, relevant sub graph is used for the sentence selection. For generic summaries, informative sentences may be chosen from each of the sub-graphs.

1.5.1.3 Discourse-based Method

Summarization problem can also be solved from a linguistic point of view, for for instance exploiting discourse relations. Mann and Thompson [43] in 1988 proposed Rhetorical Structure Theory (RST), served as a basis for the summarization approach developed in Marcu [44], extending the rhetorical relations, and using this kind of discourse representation (nucleus and satellite relations, depending on how relevant the information is) to determine the most important textual units in a document. Furthermore, Cristea et al. [45] in 2005 described an approach similar to RST differing from the previous ones, in the lack of relation names and the use of binary trees. This summarization approach is intended to exploit the coherence and cohesion of a document.

Cohesion and coherence are two of the main challenging issues for TS. Some approaches rely on the identification of such relations in order to improve the quality of the generated summaries. In 2008 Gonçalves et al. [46], used coreference chains to deal with referential cohesion problems that are frequent in the extractive summarization approach. A post-processing system is developed in order to rewrite referential expressions in the most possible coherent way, and it is applied after the summary is generated, obtaining considerable improvements in comparison to the original summaries. In order to guarantee the coherence of a summary, a widespread approach is to use lexical or coreference chains. However, the use of coreference chains is not novel in Text summarization. Baldwin and Morton [47] in 1998, and Azzam et al.

[48] in 1999 found the approaches first, where the main assumption is that the longest coreference chain indicates the main topic of the document, and shorter chains represent subtopics. Therefore, one possible strategy for building summaries is to select only those sentences related in the longest chain. This strategy helps to maintain the coherence of the text. A similar idea is to use *lexical chains*, which consists of determining sequences of semantic related words (for example, by concept repetition or synonymy relations). By using lexical chains, the main topics of a document can be also detected. This technique has been also widely used in summarization, and approaches like the ones described in [49] [50] [51] exploit them to produce summaries.

1.5.1.4 Cluster-based Method

Documents are normally composed with the end goal that they report distinctive topics in a steady progression in an ordered manner. They are generally divided up explicitly or implicitly into sections. It is instinctive to think that summaries should address diverse "subjects" appearing in the documents. Some summarizers comprise this perspective through clustering. Document clustering becomes essential to generate a meaningful summary, where the document collection for which summary is being produced is on totally different topics. The selection of sentence depends on the similarity of sentences to the subject of the cluster C_n. The other feature that is considered for the selection of sentence is sentence position in the document (P_i). If the sentence appears in the starting of the document, it contains the higher weight for inclusion in summary. The last feature that increases the sentence score is its similarity with the first sentence in the document from which it belongs (S_i).

1.5.2 ABSTRACTIVE SUMMARIZATION APPROACH

Abstraction method of automatic text summarization is, in difference to extractive method, based on text generation techniques. In this approach the summaries contain the words not present in the original document. Generally as of language complexity and ambiguity it is hard task for computer research to solve this task successfully. In general, the tactical flow of automatic text summarization based on abstractive summarization approach shown in Figure 1.5.

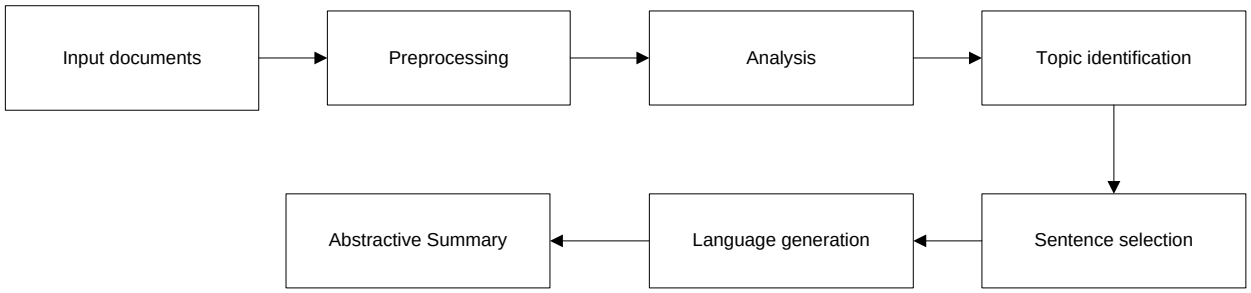


Figure 1.5. General Architecture of Abstractive Summarization Approach

Abstractive summarization involves paraphrasing the input text to produce a final and rephrased summary. From abstractive summarization, the resulting summary might contain inferred information from the given documents. Abstractive summarization includes a wide range of processing methods, also require some commonsense and domain specific text skeletons which will be filled with extracted information's [52], to fully extractive summarizations that involves whole sentence generation [53]. Abstractive summary produces important materials in new generalized form; it consists to understand the source text by the use of linguistic method to interpret [37].

Abstractive summarization approaches are of two types: first is structured based, various methods like tree based, template based, ontology based, and rule based use structured approach. Second method is semantic based which uses multimodal semantic, information item based and semantic graph based.

1.5.2.1 Structure-based Approach

Structure-based approach takes out the best vigorous information by using cognitive feature schemas such as [54] templates, extraction rules and numerous other structures similar to tree, ontology.

1.5.2.1.1 *Tree based Approach*

Tree based approach used dependency tree to represent the document content. For the outline, content selection used different algorithms as algorithmic program, which uses intuitive arrangement to select sentences for a summary. For outline generation, the approach used either

a language generator or associate degree algorithm. The approach proposed by Barzilay et al. [55], for automatically fuse similar sentences through news articles on the same topic. They used sentence generator to build a syntactic structure by mapping from concepts to words. For theme identification they used DSYNT tree, where node represents the non-auxiliary words of a sentence and the node has been connected to its direct dependents. Their approach considerably improved the quality of resultant summary.

1.5.2.1.2 *Template based Approach*

Template based approach used a template to represent a full document. GISTEXTER [56], a multi-document summarization system exploits template based method for abstractive summary generation from multiple news articles based on the output of information extraction system. In this approach, the topic of a document has been represented by the template. This approach could not handle the information about multiple documents similarities and differences.

1.5.2.1.3 *Ontology based Approach*

Ontology is a formal conceptualization of a real world and provides various semantic links between concepts; also, it can share a typical comprehension of the real world [57]. Ontology helps user and system to communicate with each other by the mutual and shared understanding of a domain [58]. Lee et al. [59] proposed a fuzzy ontology based approach for Chinese news summarization to model uncertain information and hence can better describe the domain knowledge.

1.5.2.1.4 *Rule based Approach*

In rule-based approach, the documents to be summarized are described in terms of classes and listing of features [60]. It is based on abstraction scheme [54]. For generating short well written abstractive summary, rule-based custom-designed information extraction module is used which is integrated with content selection and generation, where content selection module is used to selects the most effective candidate from those generated by the data extraction rules to answer one or more features of a class. It provides the accurate summary, but the key limitation is it consumes time as rules and patterns are written manually.

1.5.2.2 Semantic based Approach

In this approach, natural language generation system accepts linguistics designed documents as input. For processing linguistic data, this approach identifies noun phrases and verb phrases

1.5.2.2.1 *Multimodal Semantic Model Approach*

Abstractive summary has been generated from unified semantic models of a multimodal document, irrespective of the original format of the information sources [61]. This approach used linguistic model to captures concepts, occurrence of an expression, and relationships among ideas. For multimodal documents, it represents the contents as text and images.

1.5.2.2.2 *Information Item based Approach*

Information Item based approach, is the smallest element of comprehensible information in a text or a sentence [62]. By generating sentences targeted to address specific information needs, this approach permits for directly answering queries or guided topic aspects. In this approach, user information need is known and the summarization system tries to address it. Its goal is to identify all entities in the text, their properties, predicates between them, and characteristics of the predicates. It gives the rich, unambiguous, structured and short summary.

1.5.2.2.3 *Semantic Graph based Approach*

The goal of this approach is to summarize a document by constructing a linguistics graph known as Rich Semantic Graph (RSG) for the source document and generated semantic graph is reduced to more abstracted graph, and abstractive summary has been generated from the reduced graph. [63]. It uses heuristic rules to reduce the generated rich semantic graph to more reduced graph. The approach includes three phases as, the RSG Creation Phase, the Rich Semantic Graph Reduction Phase, and the Summarized Text Generation Phase.

1.5.3 DIFFERENCE BETWEEN EXTRACTIVE AND ABSTRACTIVE SUMMARIZATION

Summaries that are constructed by extracting important passages, sentences or phrases from the source document are called extracts, they are easy to adapt to larger sources and resulting incoherent summaries. In contrast, an abstract may or may not contain words in common with the document and provide more sophisticated summaries. Authors using abstraction techniques are not constrained as those using extractive ones, and can summarize a wider range of materials electively and often with smaller amount of text, they adapt well to high compression rates like Personal Digital Assistants (PDAs) and similar technologies [12]. Both of the approaches discussed so far in the chapter have their remarkable capabilities and certain limitations as well as concluded in Table 1.4.

Table 1.4: Difference between Extractive and Abstractive Summarization Approaches

Factors	Extractive Summarization	Abstractive Summarization
Definition	Formulated by extracting sentences or passages from the text.	Develop an understanding of the key concepts in a document and then those concepts are expressed in a clear natural language.
Approach used	Sentences are extracted based on the statistical analysis of discrete or miscellaneous surface level features such as frequency of word/phrase, sentence location or cue words.	To examine and interpret the text, linguistic methods are used and then a new shorter text is generated, that conveys the most significant information from the source document.
Issues	Extracted sentences usually tend to be longer than average.	System cannot summarize what their representation cannot capture.
Limitations	Compression can be lost with this approach, however, because it introduces unnecessary material.	Heavy machinery required from natural language processing (NLP), and includes grammars and lexicons for parsing and generation.

1.6 DOMAIN SPECIFIC SUMMARIZATION

In previous section we discuss the general summarization whereby the summary relevance is decided simply by the input document without concerning to its domain or the user needs [64]. Domain specific summarizations have special structure or distinctive features which should be considered by the summarizer to generate more precise information. In this section we discuss some of the works regarding domain specific text summarization.

1.6.1 MEDICAL SUMMARIZATION

In the medical field study of text summarization was found to be very useful. Summarization helps doctors to get the important information about a specific disease or information from the patient records [65]. It will also help patients who want to find information about their health problems online [66]. Moreover, there are extensive resources are present that provides access to medical information and medical-related databases like MEDLINE; it is a biomedical database that covers nearly 20 million articles. Centrifuser is an early summarization system that has been built for medical knowledge [67] [68]. It helps users by generating query-driven summaries in their search for healthcare information. Some researchers used ontology information for medical summarization [69].

1.6.2 EMAIL/BLOG SUMMARIZATION

There have also been many researches are reported in literature on email and blog summarization. In the initial research on email summarization, Nenkova and Bagga [70] developed a system to produce summaries from email threads. They produce short "overview summaries" by sentence extraction through the thread root message and its immediate follow-ups. From the root messages, sentences are extracted by finding the largest similarity nouns and verbs with the email subject. Likewise, from the follow-up emails, sentences are selected by calculating the major similarity of nouns and verbs between the root email and the follow-up emails. Newman and Blitzer [71] also address the problem of summarizing email threads. First, all the messages are clustered into group messages. In each group, sentences are scored using numerous features. Then from each group, summaries are extracted.

1.6.3 NEWS SUMMARIZATION

These days News Summarization has been identified as a crucial research area, Internet users are commonly turning to the web for news instead of going to traditional sources such as newspapers or television. Online news has become one of the major channels for Internet users to get news. An important part of online information is represented by the online news. Reading

news online offers many benefits over traditional media. News websites are daily overwhelmed with plenty of news articles, nearly all news web sites are accessible free of charge [72] for example NewsIs-Free (newsisfree.com) is a group of news sites links, which contains more than 20,600 online news sources. Many of these sources generated a large amount of online news articles and updated every day. Thus hundreds or even thousands of news articles can be found on a single topic or event. This may reason a high degree of redundancy in information provided by the set of news articles. Readers can be affected with the huge volume of news. For a reader interested in a given topic, this can cause a negative effect for online news because it becomes impractical to find and read all related news stories. This raises an unavoidable problem of how users take a quick overview of the complete story regarding a particular topic. Hence news summarization is useful for automatically generated comprehensive summarization of the news articles in a non-redundant way [9].

For building a news summarization system, it is important to study how journalists write news stories. Traditionally, they are used inverse pyramid structure [73]. Usually articles are started with a broad overview of the situation or event, later followed by the finer details of the story. To the extent that writers follow this arrangement, it can be exploited by the summarizer [72]. There are typically many articles on the same event. Summarization systems need to produce a concise and fluent summary conveying the main information in the input and help readers in determining if they want to access and read the full articles as well as allow them to get the idea of the reported event by reading the summary only [74]. On the other hand Summarization is an ideal solution to provide condensed, informative document reorganization for faster and better representation of news evolution.

1.7 NEED OF NEWS SUMMARIZATION

In online news, probably the most obvious gains are accessibility and recency, as users can read stories instantly once they are published, from any place in the world. News websites have been around for more than two decades. However, until recently, they still shared a similar, manual publishing process with their printed counterparts.

Web news service providers clusters news stories from different news websites and present them to the user. The various news stories for one category, most likely on an identical topic, have major overlaps in their contents, while some of them are unique compared to rest of the stories. Therefore there is a need to have a method which will provide a single, preferably short and informative article summary that gives the user a consolidated story regarding particular news topic.

Major online news outlets serve tens of millions of monthly visitors. When a new story is published on the main page, its instant exposure is huge. Without the delay and practical limitations of print, the number of potential stories is virtually unlimited. This can easily lead to informational overload, since human attention span is relatively constant. Therefore, the question is: how to pick the most informative and engaging stories? This leads to the need of summarization to pick the most important news on the same event from different news websites and summarize them to save the readers time to read all the news articles.

1.8 NEWS FILTERING

The rapid growth of the web makes it urgent for efficient instant online document filtering. Web page filtering is used to let users to see only those portions of a page that are useful in summarization. News web page summarization face the two main issues: the first one is how to locate relevant documents (URL's) on the web and the second one is how to filter out irrelevant documents from a set of documents collected from the web. In our work we address the second issue. In our work for filtering stage we used the web information (mainly content) extractor, which retrieves the news webpages title and news content by using Tag Tree explain in detail in chapter 4.

Content extraction is widely used in web for extracting relevant content which is useful in many areas like summarization. A web content extraction system automatically and repeatedly extracts data from web pages with changing content and delivers the extracted data to other applications in this work for the summarization.

The revolution of the Web generates various information sources published as HTML pages on the Internet. Though, web contains many redundant pages such as mirror sites or identical pages with different URL. The news web page encloses the category information, advertisements, news content, latest headings, related news etc. regarding these content parts as content blocks, all blocks apart from the news content block are identical in all news pages. Content extraction extract data from unstructured texts that are written in natural language and produces structured data ready for post-processing, which is crucial for summarization.

1.9 TEXT CLASSIFICATION

In news web page filtering and summarization, the task of news web page classification has remained in sharp focus since long. A news web page classification phase classifies a news web page from a non-news web page. Prior knowledge of correct news page allows us to narrow our content extraction task considerably. Therefore, news web page classification is considered as most important task in the news filtering and summarization.

Text classification (TC) is an area where classification algorithms are applied on text documents. In the process of TC, a set of input document is divided into two or more classes where each document can be said to belong to one or multiple classes based on the contents of the documents [75]. Typically, these classes are labeled by humans. TC dates back to the early ’60s, but only in the early ’90s it became a major subfield of the information systems discipline. It plays an important role in the field of natural language processing or other text-based knowledge applications, especially with the recent explosion of readily available text data such as electronic news articles, and digital libraries.

There are two sorts of ways to deal with text classification: rule based and machine learning based approaches. In the rule based approach classification rules are defined manually and documents are classified based on rules. Rule based approach gives good results when a number of rules are small and written by the specialists; otherwise maintenance of rules becomes more difficult as the number of rules increases [75] [76] [77]. Sometimes rules conflicts each other and have to be reconstructed when the target domain changes. To overcome these limitations machine learning approach is used in text classification. Machine learning helps us to

categorize the documents automatically [78]. In this approach rules or equations are defined automatically using sample labeled documents. It classifies text documents automatically and gives a high analytical performance. It is domain independent. In this approach a classifier observing the characteristic of a set of documents; from these characteristics classifiers should decide a new unknown document comes under which category. A classic example of text classification is categorizing news articles into topics such as politics and sports. Our objective here is to show that text classification techniques allow a much more refined analysis of the impact of news web pages filtering and summarization. This approach provides excellent accuracy, reduces labor, and ensures conservative use of resources.

Text classification is easily identified by three paradigms: binary case, multi- class case and multi- label case.

- **Binary case:** A sample belongs to exactly one of two given classes.
- **Multi-class case:** A sample belongs to just one class of a set of n classes.
- **Multi-label case:** A sample belongs to several classes at the same time therefore classes may overlap through documents.

1.9.1 PROCESS OF TEXT CLASSIFICATION

The process of text classification consists of following phases:

a) Pre-Processing

The pre-processing phase used to present the text documents into clear word format. In this phase three main steps tokenization, stop word removal and stemming words are performed. In tokenization, document is considering as a string and then partitioned into a list of tokens. Stop word removal, removes the stop words like “the”, “a”, “and”, etc because they occur frequently in the document and shows less significance. Stemming word converts different word form into similar canonical form by using stemming algorithm. After these steps document is prepared for the next phase.

b) **Document indexing**

This phase takes the training, validation and test documents as input, and outputs internal representation for them.

c) **Learning Classifier**

This phase takes the representations of the training and validation documents as input and outputs a classifier.

d) **Evaluating classifier**

This phase takes the results of the classification of test set as input and accomplished by evaluation techniques belonging to both the information retrieval (IR) and the machine learning tradition.

The process diagram of text classification is shown in Figure 1.6.

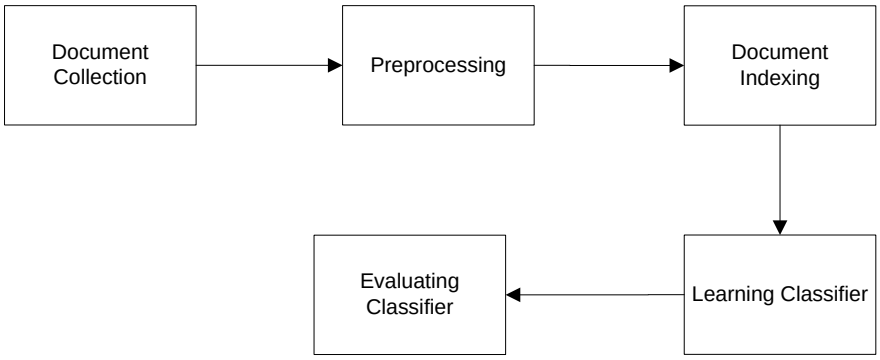


Figure 1.6. Text Classification Process

1.9.2 TEXT CLASSIFICATION APPROACHES

Text classification approaches are used extensively to solve real-world challenges. The general taxonomy of text classification approaches has been shown in Figure 1.7. Text classification approaches are classified as:

- **Supervised approach**

  These approaches use machine-learning techniques to learn a classifier from labeled training sets. Supervised approach is based on annotated corpora.

- **Unsupervised approach**

  Unsupervised approach is completely relies on external information and does not use sense tagged data (training data).

- **Semi-supervised approach**

Semi-supervised approach has both sense labeled and unlabeled data that employed different proportions to learn a classifier.

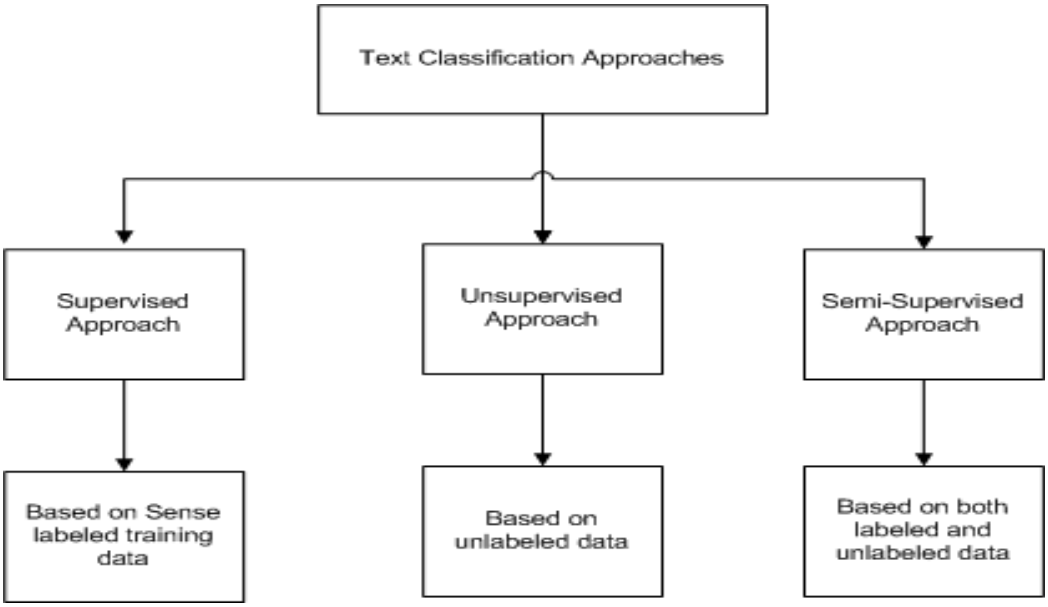


Figure 1.7. Text Classification Approaches

Supervised approach undoubtedly performs better than other approaches [79]. If we perform an unsupervised classification, then there is no idea about the possible final classes. Thus we used unsupervised learning approach to find possible groups or classes, called document clustering. Semi-supervised approach uses small amount of labeled data with a large amount of unlabeled data. According to Lu [80], Knowing unlabeled distribution for identifying classes does not help without making appropriate assumption about the relationship between labeled and unlabeled distribution. This is a limitation of semi-supervised approach. If we have predefined class or classes then supervised approach is used to achieve the solution of the problem.

1.9.2.1 Supervised Approach

Text classification problems can be solved by applying supervised learning algorithms where the learning process is supervised by the knowledge of the classes and of the concerned training instances and results is deduced from a training set [81]. For the classification of document, classifier needs to learn some basic knowledge. Therefore, the input objects are

divided into training and testing data. In Training datasets, documents are already labeled. It is initially categorized by the experts while in testing datasets, documents are unlabeled. Supervised learning algorithms analyze from the already labeled training dataset and learn the knowledge and apply this on the testing dataset for accurately predicting the class label.

Some important algorithms of supervised machine learning are Naïve Bayesian, Support vector Machine, decision tree, K-Nearest Neighbour.

1.9.2.1.1 *Naïve Bayes Algorithm*

This algorithm comes under probabilistic approach. It is a statistical method and estimations a set of probabilistic parameters that intended the restrictive and combined probability distribution of classifications and context. It based on Bayes rule, and relies on very simple representation of documents [82]. A naïve Bayes classifier is built by choosing the best sense for an input vector amount to choosing the most probable sense. It calculates the conditional probability of each sense Si of a word w where fj are featured in the context. Naïve Bayes classifier finds class that maximizes the formula by choosing S as the most appropriate sense in context. Assuming that all features are independent of each other than algorithm can be stated as in (1).

S= argmax P (Si|f1,…., fm) = argmax P(f1,……, fm | Si) P(Si) / P (f1,……, fm) (1)

Where P(Si) and P(Si/fi) are estimated during training process using relative frequencies. Short computational time for training is the main advantage of the Naïve Bayes classifier.

1.9.2.1.2 *Support Vector Machine (SVM) Algorithm*

SVM is a learning algorithm for classification used for any classification problem (text classification as one example). SVM is a binary classifier i.e. it takes care of classification problem of two classes. It is a large margin classifier, based on the vector space. SVM is utilized to find a decision boundary between two classes that is maximally distant from any point in the training data and the distance of the decision surface to the nearest data point is known as the classifier's edge. Those points which describe the position of the separator are known as support vectors. For good classification decisions margins must be maximized. SVM separates the

training set with two ways, the first way of document separation is linear separation and when adds complication then usage kernel functions.

For classification, once the hyperplane has been generated new points are mapped into a feature space by using kernel function. Some examples of the kernels are: linear kernel, Gaussian kernel, Exponential kernel, polynomial kernel, Hybrid kernel, etc. Selection of an appropriate kernel function is an important task. It is common practice to select the best one. For this reason SVM shows the drawback of the low speed of training. Fast Training of Support Vector Machines using Sequential Minimal Optimization (SMO) is implemented by Platt in 1998 [83].

1.9.2.1.3 *Decision Tree Algorithm*

Decision tree is a prediction based model [84]. It is characterized by a hierarchy of rule in the form of a tree, and utilized basically in classification problem. Training data is used to construct the decision tree. A tree is constructed to model classification problem with this technique. Exactly when a new sample is classified, the list of rules is checked and the standard that matches is applied first. Internal nodes of a tree represent a feature on which test is accompanied, topmost internal node is called the root node. Each branch represents a feature value that is the result of the test performs on the internal node (feature). Each leaf node represents a sense or a class. Several specific versions of decision tree are available such as J48, ID3, C 4.5 and CART (Classification and regression tree) etc. The advantage of the decision tree is that it is simple to understand and interpret. It performs well in huge amount of data.

1.9.2.1.4 *K – Nearest Neighbour Algorithm (KNN)*

This algorithm performs the classification, by contrasting a given test sample with training samples which are related to it. Whenever a new point is found in the classification, its k-nearest neighbor is found first from the training information. KNN determines the label of unknown vector by utilizing its K nearest neighbors. It contains numerous three principal components: firstly a set of labeled approach, for instance a set of stored records, Second a distance or comparability metric to compute distance between objects and last the value of k, those numbers for claiming nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled object is computed and its k-nearest neighbors are identified. Class

labels of the objects are determined by the class labels of their nearest neighbors. Closeness or similarity of the senses is measured by the distance metrics such as Euclidean distance.

1.10 CHALLENGES IN NEWS SUMMARIZATION

Online news reading is beneficial over traditional media. Now days thousands of news sources are available and almost all news websites are free of charge. News summarization systems help the readers to read accurately and concise summary of a particular topic rather than reading a full document. Along with these benefits there are some challenges of summarization.

- Commonly, humans are very capable summarizers, because we have an incredible talent to read an article entirely and then write a summary highlighting its main points. However, machine generating summarization, posts a rather challenging problem to computer scientist, since computers lack human knowledge and language capability; it makes news summarization difficult and non-trivial task.

- One of the major challenges in extractive summarization is summary cohesion. Sentences are usually extracted and just concatenated to each other, it is quite common that there exist no smooth transition between topics in different sentences, to overcome these cohesion problems lexical chain is introduced.

- The another challenge in performing summarization is how to identify what is the most relevant part of the article and how it can be identify that the relevant part of the article to obtain the most relevant in summary.

- The other challenge in summarization is focused on finding keyphrases and these were usually based on a term or a phrase appearing in the article. If a sentence contains identified keyphrase but top of that some other information, which is not related to the previous sentences while has high quality information and might be found to be more relevant than other.

- Another challenge is the summarizers evaluation; Evaluating the quality of a summary is identified as a difficult problem, mainly because there is no clear "ideal" summary. If someone belief that the summary is certainly a reliable substitute for the source, they must be confident that it does in fact reflect the relevant content from the source. The summary readability in terms of grammar and coherence has difficult to be evaluated.

- Therefore, methods for generating and evaluating summaries must complement each other.

- In multi-document summarization a challenging aspect is that content and writing style may vary significantly from source to source. The differences in style can make it challenging to detect how two documents relate.

- In a multi-document summarization, there is an evidence of duplication or redundancy in the summary. This is because information in document A will also appear in document B and this even makes such information good candidate for summary. Only one of such sentence should remain in the summary but perfectly removed redundancy is difficult because of lack semantic training.

1.11 MOTIVATIONS AND RESEARCH GAP

News summarization is immensely helpful for trying to figure out whether or not a full news article meets the reader's need and is worth reading the full article for further information. News summarization system summarizes various news articles on the same topic and help readers to read the precise summary which contain the main information of the articles. The key to summarization is to reduce the length of the article while retaining its main points and overall meaning.

A good number of researches have been done and is under process in the field of news summarization systems. Some of the important areas for which summarization efforts have been made are web summarization, email summarization, scientific articles summarization, video summarization, news summarization etc., however, there is rarely any method available that is

used for all kind of domains. This is due to the fact that a particular method that works on one domain may not work on other domain effectively. Many summarizers create summary by extracting salient sentences from the input articles.

Most of the earlier works were based on the single-document summarization. In the single-document summarization, approaches are based on sentence extraction from documents [87] [95]. It is specific approach for single-document summarization. Most single-document summarization systems used a simple way for summary generation by taking the first sentence from every paragraph and place them along in their original order. Later on, due to the presence of numerous sources carrying the same information cause difficulties for end users of news providers; they must read the same information over and over again. Therefore recent efforts have focused towards multi-document summarization [11] [46]. For the multi-document summarization valuable strategies are required to merge information stored in different documents. This would typically mean that certain operations need to be taken below the sentence level, probably involving keyphrase match, matching terms, sentence position, and sentence length. Therefore, multi-document summarization may effectively addresses the issues by using a metric for reducing redundancy and maximizing diversity in the selected articles and generates shorter summary containing the main points of the original documents. It resolves the problem of information overload.

From the earlier researches, we found certain issues such as, the selection of sentences, this issue arises when the diverse contents are selected from different news articles and some of the sentences are not strongly related to each other. According to Lee at al. [59], summary many a times contains sentences that are not strongly related to each other. This can be addressed by selecting a suitable threshold to generate the sentence set. Thus the selection of suitable threshold is the one of our problem. The next issue is the ordering of sentences in the summary. According to Radev et al. [72] sentences are taken from various source articles and put together to form a summary, so the resultant summaries sometimes do not seem to flow as much as they should, and they should be difficult to understand. Therefore correct ordering of sentences is required. Another issue with news summarization system is handling of large feature sets, as the complexity of tuning of weight experimentally increases exponentially. Thus there is a need of high performance systems with more effective features. The other important issue on which not

much attention has been paid so far is the combination of features. The selection of suitable content for news summarization has also been an issue as reported in literature. We observed that summary combination needs improvements in the process of content selection, as the diverse content get selected from different news sources.

Our motivation to pursue the present research is based on the works carried by different researchers underlining the issues reported in the literature as discussed above towards and efficient and effective news summarization system. The detailed review of available literature as discussed in chapter 2 also reveals these facts.

More precisely, in the proposed work we have worked on addressing the above mentioned issues, keeping an objective is to design a news summarization system by correct content selection and explore how the optimal weight can be obtained automatically by selecting minimum feature set and reduce redundancy.

1.12 OBJECTIVE OF RESEARCH

The main focus of news summarization system is to express the important ideas of news stories by eliminating less important and redundant information. News summarization is immensely helpful for trying to figure out whether or not a lengthy news article meets the user's needs and is worth reading for further information.

In order to carry out the proposed research work we have fixed the following objectives:

1. **Study different existing tools and techniques of supervised text classification for news filtering and summarization on the web:** As the growing number of news stories published online there is a need for news summarization systems. News summarization used to reduce the length and detail of article while retaining its main points and overall meaning. We study the existing work related to news filtering and summarization of many researchers in literature reviews (Chapter 2).
2. **Classifying news content on the web:** Classification algorithm identify web news page from web by selecting important attributes. It is based on the combination of content, structure, and URL attributes and Naïve Bayes algorithm is used to distinguish news

articles from non-news articles. We discuss the detailed process of news web classification in chapter 3.

3. **Filtering the web pages after classification:** News web page classification (discussed in Chapter 3) is an important phase in news summarization. Classified news web pages are further used in filtering. It provides high quality news content for analyzing. For filtering content extraction approach is used which tokenize HTML pages and construct the Tag Tree for pattern matching and filtering. Detailed approach is discussed in chapter 4. Extracted content is further used for the keyphrase extraction (Chapter 5).
4. **Keyphrases extraction:** keyphrase extraction captures the main phrases of the news web pages (discussed in Chapter 5). Keyphrase extraction approach identifies candidate phrase from the documents and chooses those candidate keyphrase having highest weight score. Weight formula combines the feature set that includes TF*IDF, phrase distance in documents and lexical chain that is based on WordNet to represent semantic relations between words. Extracted keyphrases are used for sentence extraction in summarization (chapter 6).
5. **News Summarization:** News summarization phase used similarity measure and sentence selection and ranking. Keyphrases (discussed in chapter 5) plays an important role in sentence selection. It shows favorable results compared to other news summarizer. We discussed the proposed approach of news summarization in chapter 6 and results and discussion are mentioned in chapter 7.

1.13 ORGANIZATION OF THESIS

The study has been undertaken with the following chapter scheme:

Chapter1: Introduction introduces automatic text summarization and provides details about types of summarization, news summarization and the approaches of summarization. The issues presented in summarization system. At the end, we have provided a formal description about the research objective that intends to be addressed in this thesis work along with the rational of research.

Chapter2: Literature Review discusses the existing work in the field of news filtering and summarization system. The chapter gives the detailed literature of existing news summarization systems. The focus of the research is to outline previous research on news summarization, with

particular emphasis on the content extraction and keyphrase extraction for summarization. It also looks at related techniques to news summarization and show how research in these areas can profitably be used for news summarization systems.

Chapter 3: **News Web Page Classification** shows a news web page classification phase for classifying a news web page from a non-news web page. Classification is based on the three attributes Content, structure and URL using Naïve Bayes algorithm. News web page classification is considered as most important task in the news filtering and summarization and allows us to narrow our content extraction task considerably.

Chapter 4: Content Extraction from News Web Pages using Tag Tree presents a content extraction approach for news summarization. This chapter introduces an approach for extracting the main content from news web pages.

Chapter 5: Keyphrase Extraction of News Web Pages shows a keyphrase extraction approach for news summarization based on the features TF-IDF, phrase distance, and lexical chain to calculate the score of the candidate phrase for the identification of keyphrsae.

Chapter 6: System architecture of News filtering and Summarization present the overall proposed architecture of News filtering and summarization system. It provides a complete processing of sentence selection and ranking and reduces redundancy by cosine similarity measure.

Chapter 7: Result and Discussion discusses the experimental results of overall system and shown the comparison between the existing approaches and the proposed system approach.

Chapter 8: Conclusion and Future Work providing a summary of the research work undertaken, contributions of this research work, limitations, and future directions in which this work could be extended.

1.14 SUMMARY

In this chapter, we discussed introduction to automatic text summarization, types of automatic text summarization, various approaches, domain specific summarization, News summarization, motivation to research and research gap, Objectives of research. Finally, chapter concludes with brief organization of the thesis. The next chapter provides a more extensive survey of the existing literature in the field of news filtering and summarization.

Chapter 2

Literature Review

CHAPTER 2

LITERATURE REVIEW

2.1 INTRODUCTION

Previous chapter has discussed many different aspects of news summarization systems underlying general architecture, types, issues and applications. Many successful systems have been developed so far found on either abstractive summarization or extractive summarization but trade-off between accuracy of the generated summaries has always remained in sharp focus. The aim of this chapter is to take a look of prominent news summarization systems and analyze the existing techniques, approaches, ideas from the field of news filtering and summarization.

The survey tries to address the all the steps of news filtering and summarization. A considerable amount of researches have been carried out in news web page classification, content extraction, keyphrase extraction and summarization. For convenience, the entire review of literature has been grouped according to these steps.

2.2 HISTORY OF TEXT SUMMARIZATION

Research in automatic summarization started to attract the attention of the scientific community in the late fifties [3]. In recent years, a wide range of techniques and paradigms have been proposed to tackle this research field [85].

Most summarization algorithms today aim at the generation of extracts given the difficulties associated to the automatic generation of well-formed texts in arbitrary domains. It is generally accepted that there are a number of factors that determine the content to select from the source document and the type of output to produce [8]. According to Goldstein et al. 2000 [11], to produce a summary automatically is very challenging. Issues such as redundancy, temporal dimension, coreference or sentence ordering, to name a few, have to be taken into consideration especially when summarizing a set of documents (multi-document summarization). When multi-

document summaries are required for a set of documents retrieved from a search engine in response to the query, the Maximal Marginal Relevance (MMR) method [86] can be applied. In the case of generic summarization, computing similarity between sentences and the centroid of the documents to summarize has resulted in competitive summarization solutions [87] [88].

Summarizations in languages other than English are not rare [89] for Scandinavian languages and the SUMMARIST project [90] for summarization of a variety of languages including Korean and Spanish. The 2005 Multilingual Summarization Evaluation concentrated on summarization from mixed input in Arabic and English, where the challenge was to generate output from automatic translations [23]. In a multi-lingual environment, the first large scale effort for the production of summaries was the focus of a Johns Hopkins Research Workshop [91] which produced SummBank, the first cross-lingual summarization framework for research in this field.

Summarization falls naturally into two subtasks: extraction and abstraction. Due to the limitations in natural language processing technology, abstractive approaches are restricted to specific domains. In contrast, extractive approaches commonly select sentences that contain the most significant concepts in the documents. These approaches tend to be more practical. In contrast, extractive approaches commonly select sentences that contain the most significant concepts in the documents. These approaches tend to be more practical [92].

2.3 LITERATURE REVIEW ON NEWS WEB PAGE CLASSIFICATION

Classification is traditionally posed as a supervised learning problem in which a set of labeled data is used to train a classifier which can be applied to label future examples [93]. Extensive research has been carried out on web news page classification, as there exist a good set of training documents for each predefined category and numerous procedures have been applied for effective experimental results [94] [95] [96]. Overall, the different type of attributes can be used for the classification of research in this area. The three main trends in attribute types are: URL attribute, content attribute and structure attribute. URL is the more informatics sources of information for classification.

Cruz et al. [97] in 1998 introduced some type of distance functions that measure structural similarity between web documents. They look at three different methods for defining similarity based on Tag Frequency Distribution Analysis (TFDA), parametric functions and edit distance between documents.

Wong et al. [98] in 2000 introduced a labels discovery algorithm that uses the hierarchical structure to represent the relation among text data in the web documents extracted from the web. Their algorithm successfully discovers similar labels which describe the same kind of information and accurately classifies web pages.

Agrawal and Srikant [99] in 2001 used a model by the combination of documents from different taxonomies. Their model used Naïve Bayes algorithm for classification task. Some research applications like NewsDude [12] unambiguously use content-based classifiers both to choose valuable articles and to eliminate articles that appear to be excessively repetitive with previously seen articles.

Sun et al. [100] in 2002, proposed the use of support vector machine (SVM) classifier to classify web pages using both their text and context feature sets. They used WebKB dataset for their web classification method experiments and results show that the use of context features especially hyperlinks can improve the classification performance significantly.

A paper written by Daniele Riboni [101], in 2002 focused on feature selection for web page classification. Robini experimented with several feature selection techniques based on information gain, word frequency, and document frequency. The experiments were made using a large set of samples from the subcategories of the Science directory from Yahoo!4 , using a Naive Bayes' classier and a kernel perceptron.

Joshi et al [102] in 2003 produced an alternative scheme for representing the structural information of documents based on the paths contained in the corresponding tree model. They define a news family of meaningful structural similarity measure that includes partial information about parents, children and siblings. Their experimental results based on the SIGMOD XML dataset show that the good clusters of structurally similar pages are produced by that representation.

Holden and Freitas [103] in 2004, used the Ant Miner algorithm in the field of web content classification and shows that it is more effective than C 5.0. It also investigates the benefits and dangers of methods to reduce the large numbers of attributes associated with web content mining such as a Naïve WordNet preprocessing stage.

An interesting paper by Kan [104] in 2004 studies the use of URLs for web page categorization. The assumption is that the URL inherently encodes information about a page's category. In the paper, Kan describes several methods for extracting tokens from URLs. The experiments used a Support Vector Machine as the classifier, and tried using different sources for text data as a comparison to using URLs only.

Selamat and Omatu [105] in 2004, proposed a news web page classification method (WPCM) using neural network where inputs are obtained by both the principal component and class-profile based features and each news web page is represented by the term-weighting scheme. The principal component analysis (PCA) has been used to select the most relevant features for the classification when the number of unique words in the collection set is large. The final output of the PCA is combined with the features vectors from the class-profile which contains the most regular words in each class. Their experimental results show that the WPCM method provides acceptable classification accuracy with the sports news datasets.

Kan and Thi [106] in 2005, presents an approach to the segment the URL into important portions and ads components, sequential and orthographic features to model silent features. Their results show that URL features perform well on classification tasks and surpass the performance of full content and link-based approaches.

Tombros and Ali [107] in 2005 consider that the textual content contained within common HTML tags, the structural layout of pages, and the query terms contained within web pages can effect on web page similarity. Their study shows that combination of features can yield more promising results than individual factors.

Tongchim et al [108] in 2006 propose a simple yet efficient technique to mine news articles from web collection. They create a dataset by gathering information from two different

Thai newspaper websites. They explore machine learning approaches to separate news web pages from non-news web pages based on the structured features of web documents.

According to Chy et al. [109] in 2014 the task of news classification is to automatically classify news documents into predefined classes based on their content. They proposed an approach to classify news articles to specific classification. They used web crawler to extract article content and construct a full-text-RSS. They apply Naïve Bayes classifier for classification of Bangla news article content based on news code of IPTC. Their experimental results show the effectiveness of their system.

Singh et al. [110] in 2016, proposed a rough set-based feature selection method to remove the redundant and irrelevant features in order to improve the performance of classifier. They used various dataset with the various supervised learning algorithm to test their method and found that their proposed method gives better performance in comparison with other methods.

2.3.1 CONCLUSIVE FINDINGS

In the phase of news web page classification, the survey reveals that most of the researches have focused on individual attributes. Some previous researches [97] [98] [102] [107] works on structural similarity of web pages for the classification. The structural knowledge of the web pages could be used to distinguish web pages. Therefore discovering such structural attributes can lead to successful web page classification. If label could not be discovered more accurately then similar labels could be discovered with fewer errors and does not give accurate classification results. Later some researchers [100] [103] [109] proposed their work on content attributes based classification of web pages. The limitation of it is that the extracted word list is usually large and it is impractical for classification algorithm.

Some researches [104][106][104] in Classification of web pages by URL attributes show that all web pages have URL's attributes regardless of whether they exist , are accessible, have incoming links or have any text. While URLs attributes do not perform as well with typical web site entry points like domain name, as they attempt to leverage the internal path structure of the URL.

Therefore in this work, we used the combination of all the three attributes for the correct classification of news web pages from non-news web pages. Some prominent research in the area of news web classification is shown in Table 2.1.

Table 2.1: Some Prominent Research Work in News Web Page Classification

S No.	Authors	Year	Description
1	Agrawal and srikant [99]	2001	Used the combination of documents from different taxonomies in their model that used Naïve Bayes algorithm for classification task.
2	Sun et al. [100]	2002	Used both their text and context feature sets for web page classification using support vector machine (SVM) classifier.
3	Selamat et al. [105]	2004	Proposed a news web page classification method (WPCM) using neural network. For the classification relevant features has been selected from the principal component analysis (PCA).
4	Kan and Thi [106]	2005	Proposed an approach based on URL to model silent features.
5	Tongchim et al. [108]	2006	Proposed a technique to mine news articles from web collection. They explored machine learning approaches to separate news web pages from non-news web pages based on the structured features of web documents.
6	Chy et al. [109]	2014	Proposed an approach for Bangla news articles classification. They used web crawler to extract article content and construct a full-text-RSS and then apply Naïve Bayes classifier for classification.
7	Singh et al. [110]	2016	Proposed a rough set-based feature selection method for removing redundant and irrelevant features in order to improve the classifier performance.

2.4 LITERATURE REVIEW ON CONTENT EXTRACTION

Plenty of work has been reported in literature for content extraction from news web pages including standard and independently techniques. Now a day many researchers paying attention in the field of content extraction from news web pages. Among the variety of work reported in literature, we reviewed the various other techniques that are attempting to solve the similar problem.

Muslea et al.[111] in 1999 presents an approach to wrapper induction which is based on the idea of hierarchical information extraction. They introduce an inductive algorithm, STALKER that generates high accuracy extraction rules based on user-labeled training

examples. Their experimental results show that STALKER does significantly better than other approaches.

Lin and Ho [112] in 2002 propose InfoDiscoverer system for the identification of informative contents from a web page. Their system partition the web page into several content blocks according to HTML tag <Table>. Their proposed method dynamically select the entropy-threshold that partition the block into either informative or redundant. Their experimental results show the value of precision and recall greater than 0.956.

Knoblock et al [113] in 2003developed a set of tools for extracting data from web sites and transforming it into a structured data format such as XML. Their approach automatically detecting the breakage of wrapper and repairing them capitalizes on the regular structure of the extracted fields themselves. Their technology learns highly accurate extraction rules and wrapper is verified by the correct extraction of data.

Reis et al. [114] in 2004 present a domain oriented approach to web data extraction. Their approach of data extraction from web pages is based on the structure analysis of the target web pages. The structure of the web page can be described by a DOM tree, since they introduced a new algorithm RTDM for calculating the edit distance between two given trees and solve the problem of structure- based page classification, extractor generation and data labeling. Their experiment results show that RTDM correctly extracting 87.71% of news from a dataset of 4088 pages.

Gupta et al. [115] in 2005 developed a framework that use DOM trees and a W3C specified interface that allows programs to dynamically access the structure of a document. Their approach is implemented in a publicly available Web proxy for the content extraction from HTML web pages.

In 2007, Banko et al. [116] introduces a fully implemented, highly scalable OIE system TEXTRUNNER, which has the ability to extracts large amount of high quality data from a nine million web page corpus. In this system tuples are assigned a probability and indexed to support efficient extraction and exploration via user queries. They compare TEXTRUNNER with KNOWITALL, and achieve an error reduction of 33%.

Gibson et al. [117] in 2007 used machine learning methods to identify the content of a news web page. They identify correctly the portion of the content of web pages from 80-97% of the time. Their experimental result shows that the document level accuracy of their model is 80% and also shows the high value of precision and recall for content block identification.

Ziegler et al. [118] in 2007 present an approach that works in a fully- automated fashion, classifying text blocks in html pages using distilling linguistic and structural features, having a Particle Swarm Optimizer (PSO) learn feature thresholds for optimal classification performance. Their approach shows good results for hundreds of news pages from popular media in five languages and exhibits an accuracy that comes close to human judgment.

Prasad and Paepcke [119] in 2008 developed a heuristic technique for the extraction of main contents of a news web page. They construct DOM tree of the web page and scored the nodes according to amount of text and number of links it contains. There method is site- independent and does not used any language based features. The performance of their algorithm achieved 97% precision and 98% recall which is slightly below the baseline system but requires significantly less computational speed; the processing speed of algorithm is less than 15 ms per page.

In 2009, Louvan [120] propose an approach for the extraction of main content from web documents by using the combination of machine learning and heuristics approaches. Their results show that the combination of classification task and LBS namely CS+LBS gives best performance for blogs and news datasets.

Ji et al. [121] in 2010, proposed web information extraction method that is based on Tag tree template and efficiently extract meaningful information including records and data schema. They describe a web page by HTML Tag tree and both HTML tag and text are treated as tree nodes. Their result shows the effective extraction of meaningful information.

Guo et al.[122] in 2010 propose an approach called ECON for content extraction from news web pages. The approach finds a snippet- node which is used to wrap the news content and then backtracks from the snippet-node until a summary node is found, and then the summary node wrapped the entire news content. In this approach backtracking removes the noise. Their experimental results show that ECON can achieve high accuracy of 90 % for scalable extraction.

Sleiman and Corchuelo [123] in 2013, propose an unsupervised information extraction approach TEX, works on the idea that two or more web documents generated by the same server side template and removes shared token sequences among web documents until finding the relevant information that should be extracted from them. TEX working on malformed web documents and reduces extraction time by not converting HTML code into XHTML and DOM trees. There technique achieves a very high precision and recall value of approximately 100%.

Kaddu and Kulkarni [124] in 2016, Present a hybrid approach for the extraction of web page main contents. Their approach is based on the combination of automatic extraction and manual hand crafted rule techniques. They generate rules by machine learning method, by using that rules relevant content from web pages are extracted. Their work generates effective rules and achieves automaticity and efficiency. Their results show that retrieval accuracy depends on the size of the dataset. For a lower number of files it shows 28-30% accuracy and for higher number of files it shows 90% retrieval accuracy.

Pettersson et al. [125] in 2016 presents HistSearch tool for automatic information extraction from historic text. They present the outcome of collaboration between the field of computational linguistics and history, which resulting a graphical user interface for information extraction from historical text. They describe the workflow of the system, based on the spelling normalization using advanced taggers and parsers available for the standard modern language. A prototypical graphical user interface used by the historians and a Manual evaluation of the tool performed by the actual users. Their results show that spelling normalization is successful for the task of tagging and lemmatization.

2.4.1. CONCLUSIVE FINDINGS

Researches have shown that earliest extraction methods rely on human to encode the template in a program called wrapper to parse HTML on web sites [113] [116]. Content extraction that depends on extraction rules do not usually adapt well to changes to the web and it is uncommon that changes may invalidate the existing extraction rules. Like our work, Xi [121] parse HTML into Tag tree to extract template and use repeating patterns and heuristic rules to find schema in the exclusive field of template, but not to explore the semantic relation. The

limitation of their approach is to find repeating patterns and summarize the heuristic rules in exclusive content out of template to specify the schema of record in complex situation. Therefore this work does not rely on extraction rules like previous approaches and use the concept of tokenization of HTML page; these tokens construct the Tag tree and generated a template from each web page and discover matching patterns and multiple sequence alignment and apply filtering algorithm to filter out irrelevant content. Table 2.2 shows some research work in brief in the area of content extraction.

Table 2.2: Some Prominent Research Work in Content Extraction

S No.	Authors	Year	Description
1	Muslea et al. [111]	1999	Introduce an inductive algorithm, STALKER that generates high accuracy extraction rules based on user-labeled training examples.
2	Lin and Ho [112]	2002	Present InfoDiscoverer system for the identification of informative contents from a web page.
3	Knoblock et al [113]	2003	Automatically detecting the breakage of wrapper and repairing them capitalizes on the regular structure of the extracted fields themselves.
4	Reis et al [114]	2004	Present a domain oriented approach to web data extraction based on the structure analysis of the target web pages.
5	Gupta et al [115]	2005	Present a framework that use DOM trees and a W3C specified interface that allows programs to dynamically access the structure of a document and extract content from HTML web pages.
6	Banko et al [116]	2007	Present a highly scalable OIE system TEXTRUNNER, which has the ability to extracts large amount of high quality data from a nine million web page corpus.
7	Prasad J. et al. [119]	2008	Present a heuristic technique for the extraction of main contents of a news web page.
8	Louvan [120]	2009	Present an approach for the extraction of main content from web documents by using the combination of machine learning and heuristics approaches
9	Guo et al. [122]	2010	Propose an approach called ECON for content extraction based on the snippet-node to wrap the news content.
10	Sleiman and Corchuelo [123]	2013	Present an approach TEX, working on malformed web documents and reduces extraction time by not converting HTML code into XHTML and DOM trees and finds the relevant information that should be extracted from them.
11	Kaddu and Kulkarni [124]	2016	Present a hybrid approach for the extraction of web page main contents based on the combination of automatic extraction and manual hand crafted rule techniques.

2.5 LITERATURE REVIEW ON KEYPHRASE EXTRACTION

In the previous works authors have suggested that document keyphrase can be useful in many areas as information retrieval and summarization.

Chien [126] in 1999 developed a keyphrase extraction system depends on PAT Tree based adaptive approach for Chinese and other Asian languages. Their approach reduces the reliance on rigid lexicon and sophisticated word segmentation, and compared with conventional statistics-based approaches, it can handle phrases composed of high-frequency words regardless of phrase length and used the idea of internet utilization. Their proposed approach has been successfully used in several information retrieval applications, such as automatic term suggestion, domain-specific lexicon construction, book indexing and document classification.

Turney [127] in 1999 developed the GenEx algorithm for automatic keyphrase extraction based on supervised learning. They treat a document as a set of phrases and the learning algorithm must learn to classify as positive or negative examples of keyphrases. Their experimental results show that GenEx can generate better keyphrases than a general purpose learning algorithm C 4.5 and the non-learning algorithms that are used in commercial software (Word 97 and search 97).

Martinez - Fernández et al. [128] in 2003 focus on the description of the first prototype for AKE (Automatic Keyword Extraction) using Vector Space Model technique and linguistic resources in the OmniPaper project. It is a keyword extraction system for news characterization that uses several linguistic techniques to improve the current state of the text-based information retrieval to extract news articles keyphrases.

Wu et al. [129] in 2004 introduced KIP (Keyphrase identification program) which uses sample human keyphrsaes and then learns to identifiy additional news keyphrases. KIP mines noun phrases from documents and score will be allocated to each noun phrases. Depending on the weights the words that have higher score than the threshold will be selected as keyphrases.

Witten et al. [130] in 2005 describe KEA algorithm, based on Naïve Bayes classifier automatically extracts keyphrases from text. This algorithm recognizes candidate keyphrases using lexical methods and computes feature values for each candidate by using machine learning algorithm and analyze which candidates are noble keyphrases.

Wang et al. [131] in 2006 proposed in their paper Neural Network based keyphrase extraction method. For keyphrase extraction they used the features terms frequency and inverse document frequency, whether to appear in the title or headings of the given document, and its frequency appearing in the paragraphs of the given document.

Lui et al.[132] in 2007 presents a domain independent keyphrase extraction algorithm, which distinguish keyphrases from non–keyphrases by using statistical and computational linguistics techniques combination, a new attribute set and a new machine learning method; and shown that it perform well than other keyphrase extraction methods.

Wang et al [133] in 2008 propose a system for automatic online news topic keyphrase extraction where News stories are organized into topics. Firstly candidates keywords are extracted from the single news stories and then filtered with topic information. After that a phrase identification process combines keywords into phrases using position information. Finally, the phrases are ranked and top candidate keywords are selected as topic keyphrases. Experimental results show that their system performs effectively with 70.61% precision and 67.94% recalls.

Xie et al. [134] in 2010 proposes an approach which acquires semantic features within phrases from a single document. Semantic relatedness degrees between phrases are computed using word co-occurrence information in the document, and the document is represented as a relatedness graph. Keyphrases are extracted based on the semantic relatedness features acquired from the graph. Their result demonstrates better performance than TFIDF and KEA.

Gao et al. [135] in 2011 propose a method to extract hot keyphrases from news report; their method consists a two-step process of keyphrase extraction based on TF*PDF. In their method each step uses position- weighted TF*PDF schema.

Li et al. [136] in 2013 proposed an approach by combining semantic information with KEA with the help of building lexical chain and then the length of the chain is used as a feature to construct the extraction model. A lexical chain is based on the Reget's thesaurus and the experimental results show the performance improvement of their system over the KEA.

Luo et al. [137] in 2013 propose a method to integrate the comment posts for keyphrase extraction from web news documents. They introduced several strategies to select useful comments for improving keyphrase extraction task. They used machine learning technology for comment selection; their results show that comment information significantly improves keyphrase extraction from news web pages.

After that in 2014, Li and He [138] propose a method based on the lexical chain to improve KEA keyphrase extraction, their experiments result shows improvements compare with KEA and Nguyen and Kan's method.

Hsu et al. [139] in 2015 propose subject- keyphrase concept to extract subject-keyphrases from a documents known as subject-keyphrase extraction (SKE) algorithm based on the notion of definition-use chain (DU chain) to identify subject keyphrase. Their experimental results show that SKE can successfully identify the subject-keyphrases to effectively capture the main idea of a document.

Duwairi et al. [140] in 2016 presents a framework for keyphrase extraction from Arabic news documents based on the KEA system. It based on supervised learning particularly Naïve Bayes algorithm for keyphrase extraction. They compute the two probabilities for keyphrase extraction the first probability of being a keyphrase and second not being a keyphrase. The final set of keyphrases is chosen from the set of phrases that have high probability of being a keyphases.

2.5.1. CONCLUSIVE FINDINGS

Most existing work on keyphrase extraction uses only the internal information of a document. The simplest approach for keyphrase extraction in a document is to use a frequency criterion or TF-IDF model [36]. This method was generally found to give poor performance [141]. Another important clue for keyphrase extraction is the phrase location in the document. Kea [130] and GenEx [127] used this clue for their research. GenEx is computationally expensive in training time while Kea is much faster than GenEx, but it does not use parser to detect phrases, all possible word sequences up to three words are treated as possible phrases.

In this work, we only extract keyphrases and all nouns in the document are treated as candidate phrases. We used the idea of Ercan G [142] to use the lexical chains in keyphrase extraction. We also used the combination of TF-IDF and phrase distance with lexical chain for better results. The Table 2.3 describes the journey of this area in short.

Table 2.3: Some Prominent Research Work in Keyphrase Extraction

S. No.	Authors	Year	Description
1	Chien [126]	1999	Developed a PAT-Tree based adaptive keyphrase extraction system for Chinese and other Asian languages.
2	Martinez et al. [128]	2003	Present AKE (Automatic Keyword Extraction), it is a keyword extraction system which is used to extract news articles keywords.
3	Wu et al. [129]	2004	Introduced KIP (Keyphrase identification program) which uses sample human keyphrsaes and then learns to identifiy additional news keyphrases.
4	Witten et al. [130]	2005	Present KEA algorithm, based on Naïve Bayes classifier automatically extracts keyphrases from text.
5	Wang et al. [131]	2006	Proposed Neural Network based keyphrase extraction method.
6	Lui [132]	2007	Presents a domain independent keyphrase extraction algorithm, which distinguishes keyphrases from non–keyphrases by using statistical and computational linguistics techniques combination.
6	Wang et al [133]	2008	Present a system for automatic online news topic keyphrase extraction.
7	Xie et al. [134]	2010	Present an approach which acquires semantic features within phrases from a single document.
8	Gao et al. [135]	2011	Present a method to extract hot keyphrases from news report; their method consists a two-step process of keyphrase extraction based on TF*PDF.
9	Luo et al. [137]	2013	Present a method to integrate the comment posts for

			keyphrase extraction from web news documents.
10	Li and He [138]	2014	Present a method based on the lexical chain to improve KEA keyphrase extraction
11	Hsu et al. [139]	2015	Present a subject- keyphrase concept to extract subject-keyphrases from a documents.
12	Duwairi et al. [140]	2016	Presents a framework for keyphrase extraction based on the KEA system. It relies on supervised learning particularly Naïve Bayes algorithm.

2.6 LITERATURE REVIEW ON NEWS SUMMARIZATION

Early work on news summarization can be dated back to 1990s when SUMMONS summarizer was created [74]. SUMMONS was designed for summarizing news articles on the single event like terrorist events. It was used the concept of template-driven message understanding system MUC-4 [143]. Firstly, system processes the full text and fills the template slots before synthesizing the summary from the extracted information.

Goldstein et al. [11] in 2000 proposed an extraction based multi-document summaries used Maximal Marginal Relevance Multi-Document (MMR-MD) metric for reducing redundancy and achieve high compression ratios. Their approach is different from other approaches as it is completely domain-independent and depends upon fast statistical processing to maximize the novelty of the information had been selected.

In 2000, Chen and Lin [144] also proposed architecture of multilingual news summarizer, including monolingual and multilingual clustering. They defined the concept of MUs for similarity measure, and presentation of summarization results. They select high frequent English translation and name transliteration is adopted to translate Chinese MUs into English. Their experimental results show that to reduce redundancy information decay strategy is helpful, and user can get all the information provided by the news sites.

In 2001, White et al. [145] proposed a system RIPTIDES, which was similar to the SUMMONS system. For the summarization it incorporates information extraction. They used natural disaster scenario templates for each text and used them as the input to summarization

system. The summarizer first merges the templates into events oriented structure and then summary sentences are selected according to the assigned scores to each sentence.

McKeown et al. [146] in 2001, developed a composite multi-document summarization system that uses different summarization approaches dependent on the type of documents in the input set. In their system, for the automatic identification of the input set of documents a router is used which is also invokes the appropriate summarization subcomponents. Their system performs well on summary content as compared to other systems; it is ranked third or fourth with different systems ranked ahead of it for each analysis.

After that in 2002, McKeown et al. [147] also developed Newsblaster, to summarize online news articles. The summarizer used the idea of MultiGen [55], which identifies common sentences from news article by using together machine learning and statistical techniques. Summaries are then produced by analyzing and fusing together the sentences.

From the understanding of news structure, Daniel et al. in 2003 [148] investigate the utility of sub-events in news topic. Their results showed that the utilization of sub-events can improve the performance.

Lee et al. [149] in 2003 proposed an ontology-based fuzzy event extraction agent for Chinese news summarization. In their work, for testing the performance of their summarization agent, they construct an experimental website at Chang Jung University. Experimental results show that their approach can effectively summarize the Chinese weather e-news retrieved from China Times website.

D'Avanzo and Magnini [150] in 2005 describe LAKE System based on keyphrase extraction methodology using linguistic features for identifying relevant terms in the document. Generated Summaries considered both the relevance and the coverage of keyphrases for a certain topic. Their experimental results show an average responsiveness and the high linguistic quality of the summaries. However their obtained results are very competitive to pyramid metric.

Another concept called fuzzy ontology was studied by Lee et al. [59] in 2005, to develop weather news summarization. Fuzzy ontology was found to be more suitable to treat domains with uncertainty.

Svore et al. [151] in 2007, proposed a new approach for automatic summarization using neural nets known as NetSum. They apply novel feature based on news search query logs and Wikipedia entities. They used the RankNet learning algorithm to train a pair-based sentence ranking and then score every sentence in the document to identify the most important sentences. They worked on single-document summarization to improve quick access to large quantities of information.

Litvak and Last [152] in 2008, proposed and compare two novel approaches supervised and unsupervised based on graph-based syntactic representation of text and web documents, which enhances the traditional vector-space model by taking into account some structural document features. Their experimental results show that the supervised classification provides the highest keyword identification accuracy, while the highest F-measure is reached with a simple degree-based ranking.

Li et al. [153] in 2010, proposed ontology enriched Multi-Document summarization (OMS) system to generate query-relevant summary for natural calamities related news and reports like disaster management. OMS reletes sentences onto a domain specific ontology. To generate summary, sentences are extracted based on the matching between node on the ontology and the user query and the sentences attached to that particular node will be selected to form summary.

Al-Hashemi [154] in 2010 used extractive methods for document summarization and presented work based on design a keyphrases extraction subsystem and many other features extracted from the documents to select the good sentences in the resultant summary. Their system gives high quality compressed summary.

Litvak, et al. [155] in 2010 introduce a language independent approach "MUSE" (Multilingual Sentence Extractor) based on the linear optimization of several sentences ranking measures using a genetic algorithm for extractive summarization. They used English and Hebrew to test their methodology. Their results show that MUSE performs better than multilingual approach TextRank in both the languages using either monolingual or bilingual corpora.

El-Haj et al. [156] in 2011, proposed an optimized generic extractive Arabic and English multi-document summarization technique, which used a translation summary machine. Their approach uses an Arabic version of the DUC-2002 dataset that they translate using Google Translate. They explore sentence level clustering for the multi-document summarization and also eliminate redundancy. Their approach use cluster size and selection model as parameters in extractive summarization process. The experimental results show that performance of their summarization system is good in comparison with other top performing systems at DUC-2002.

Galanis et al. [157] in 2012, proposed an Integer linear programming based extractive multi-document summarization method that jointly maximizes the importance of the sentences in the summary and their diversity, beyond a maximum allowed summary length. The results on widely used benchmarks show that the approach can attain better results.

Li et al. [158] in 2013, propose a bigram based supervised method for extractive document summarization and leverage the ILP method as a core component. They revise the ILP to maximize the bigram gain rather than the bigram coverage. Their method gives better results than the previous state of the art ILP systems on different TAC data. Their experimental results show that the improvement in system performance that depends on the supervised bigram estimation module that successfully gathers the important bigram and give them appropriate weights.

Yan and Wan [159] in 2014, proposed SRRank algorithm and use semantic role information to enhance the graph based ranking algorithm for multi-document summarization. Their algorithm used heterogonous ranking process to rank sentences, semantic roles and words. They use DUC datasets for the experiment and show that SRRank outperform a few baselines approaches.

Liu et al. [160] in 2015, presents Weibo-oriented Chinese news summarization system using multi-feature combination. They used extractive based summarization methods for the single-document summarization and extract the most significant sentences from the source Chinese news article by allocating a significance score to each sentence, considering the certain kinds of features and generate the short summary.

Cao et al. [161] in 2015, developed a ranking framework based on Recursive Neural network [R2N2] to rank sentences for multi document summarization. It transforms the sentence ranking task into hierarchical regression process by using recursive neural networks model. They designed an optimized sentence selection method based on the words and sentences ranking scores. They conduct experiments on DUC benchmark; experimental results show that their model achieves higher ROUGE score than the previous summarization approaches and makes much more accurate prediction than traditional support vector regression.

Liu et al [162] in 2016 Presented mover's distance metric (WMD), in conjunction with semantic aware continuous space representation of words, and has been proposed to accurately estimate the similarity degree between a pair of documents for effective use of summarization process. They investigate their approach to other state of the art approaches and show the effectiveness of their approach over other summarization frameworks.

Demirci et al. [163] in 2017 presented a multi-document summarization system (MDS) for the Turkish news articles. They generate a single-document from the multi-document news articles via RSS based on the Latent Semantic Analysis. They examine the performance of the system based on the number of sentences and the summarization rate and show that the performance falls for long texts and improved when summarization rate increased.

Other popular Internet news services like Altavista News and Google News, present clusters of related articles, allowing readers to easily find all stories on a given topic. However, these services do not produce summaries. Therefore, a reader seeking a quick topic overview

must choose between selecting a representative article to read in full or else skimming through all articles.

2.6.1. CONCLUSIVE FINDINGS

Research on summarization begins very early by Luhn in 1958 [3] and Edmundson in 1967 [4] and becomes one of the traditional topics in the natural language processing research. Previously many papers about summarization have been proposed [24] [35] [86]. Recently, it has attracted due to the increase of online news in the internet. In the past, the major research was stressed upon single-document summarization [151] [152] Recently, effort transferred to multi-document summarization [11] [46] [157] and multi-lingual summarization [144] [155] [161].

In our research work, extraction-based multi-document summarization is used for news summarization. Extraction- based methods usually involve assigning a saliency score to each sentence and then rank the sentences in the document. We combine all the previous phases and select some features discussed in detail in chapter 6 to calculate score of each sentence for ranking and ordering which is used to generate final summary.

Most News summarization systems have been analyzed extensively and obtained significant enhancement in performance. Some prominent news summarization systems are shown in Table 2.4.

Table 2.4: Some Previous News Summarization Systems

S. No.	News Summarization System	Year	Developer	Description
1	SUMMONS [74]	1995	McKeown and Radev	Summarizes a series of news articles on the same event, producing a paragraph consisting of one or more sentences.
2	SUMMARIST [24]	1998	Hovy et al.	Summarization has been achieved by the topic identification, interpretation and generation.
3	Marcu [44]	1999	Marcu	Determine the most important units in a text by using text coherence models and RST trees.
4	MEAD [164]	2000	Radev et al.	Generate summaries using cluster centroids produced by topic detection and tracking system.
6	RIPTIDES [145]	2001	White et al.	It incorporates information extraction to support

				summarization. The summarizer first merges the templates into event oriented structure and then the importance scores are assigned to each slot/sentence to select the summary sentences.
6	Columbia's Newsblaster [147]	2002	McKeown et al.	Provide news updates on a daily basis and groups news into stories on the same event and generates a summary of each event.
7	MSR-NLP summarizer [165]	2004	Vanderwende et al.	Generate summaries by extracting and merging logical forms portions by using graph-scoring algorithm and identify highly weighted nodes and relations.
5	NewsInEssence [72]	2005	Radev et al.	Searches other related stories for the single news story select by the user and produces a summary presenting the most salient information from the different sources.
9	FemSum [166]	2007	Fuentes et al.	A summary has been generated by taking into account a syntactic and semantic representation of the sentences, and used graph-representation to establish relation between candidate sentences.
10	OMS [153]	2010	Li et al	Generate query-relevant summary for natural calamities related news and repots. It relates sentences onto a domain-specific ontology.
11	SRRank [159]	2014	Yan et al.	Used semantic role information to enhance the graph based ranking algorithm for multi-document summarization.
12	Weibo-oriented Chinese news summarization system [160]	2015	Liu et al.	Used extractive single-document summarization methods and involved multi-feature combination for sentence extraction to generate a summary for the Chinese news articles.
13	MDS system for Turkish News [163]	2017	Demirci et al.	Introduced a muti-document summarization system (MDS) for Turkish news based on the Latent Semantic Analysis and produced summary to show the main idea of the entire article.

2.7 HIGHLIGHTS OF LITERATURE REVIEW

This chapter surveys the phases of news filtering and summarization systems. Survey findings show that much work has been done in each phase but also shows some limitations. In order to know the need for further research in each phase has been highlighted in this chapter as follows:

- In the news web page classification phase we review the previous research work on the individual attributes using different classification algorithms and found that combination of attributes gives better results than individual ones.

- In the content extraction phase, according to previous research work, extraction rules usually does not adapt well and invalidate to changes in the web. Therefore the concept of tokenization and construct the tag tree to discover matching patterns and apply filtering algorithm to filter out irrelevant content shows better results for content extraction.
- In the keyphrase extraction phase, previous researches show that TF-IDF and phrase location was the simplest approaches for keyphrase extraction. These methods were generally found to give poor performance. Therefore for extracting keyphrases, we used lexical chains with the combination of TF-IDF and phrase location for better results.
- Finally in the news summarization phase, previous researches are focused on single – document summarization then transferred to multi-document and multi-lingual summarization. In this work we combine all the above three phases and extraction-based method is used to calculate saliency score of each sentence for sentence ranking and ordering which produce the final summary.

2.8 SUMMARY

The chapter provides the detailed overview of the research work done in the area of news web page classification, content extraction, keyphrase extraction and summarization. Our literature review indicate that many significant researches has been done on each phases of news web page filtering and summarization. But they raised various issues and offer researchers to carry more in depth analysis and propose approaches to improve the performance of such systems to match end user expectations.

Chapter 3

News Web Page Classification

CHAPTER 3

NEWS WEB PAGE CLASSIFICATION

3.1 INTRODUCTION

In news web page filtering and summarization, the task of news web page classification has remained in sharp focus since long. A news web page classification phase classifies a news web page from a non-news web page. Prior knowledge of correct news page allows narrowing content extraction task considerably. Therefore, news web page classification is considered as most important task in the news filtering and summarization. In this chapter, we present an approach based on three attributes content, structure and URL using Naïve Bayes classifier.

The task of news web page classification is to automatically classify news web pages into predefined classes based on their attributes. Classification shows an important part in various information retrieval tasks. The web is very diverse in nature, and no rules are there on how to build HTML pages and how to state the entire structure of the web pages [167]. Thus automatic news web page classification becomes an important task. News web page classification technique uses a variety of information to classify a target page. The vital notion for news web page classification is the similarity measurement among web documents. Similarity analysis and classification can be done on attributes drawn from news web documents.

3.2 WEB PAGE CLASSIFICATION

Classification is considered as a supervised learning problem in which a classifier is trained on a set of data labeled with predefined categories and then applied to label future examples [81]. Based on the number of classes in a problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two classes whereas multi-class classification deals more than two classes [95]. It plays a fundamental role in a number of essential tasks on information retrieval and summarization.

Even though web page classification is similar to text classification, web pages have some characteristics not present in textual documents. Indeed, web pages have an underlying structure in HTML [101]. Web page classification process is to predict the category of the document by analyzing the distribution of words involved. Here, by analyzing web pages we can see that the contents of web pages and the words within them have inevitable connection, which means the same category of web pages must contain a large number of same words, but to the words in different category of web pages, there must be few in common. They contain noisy contents such as advertising panel or contact information. They are composed of many parts that can talk about different but related subjects and are linked to each other either by hyperlinks or by users' intuitive judgments [168]. All those characteristics incur a bias for the classifier and lead to misclassification of web pages. Thus, web pages related to a target web page can help highlight its topic and correct the class initially assigned by a classifier. A classifier is usually evaluated with regard to how accurately it can label unseen instances. The accuracy of the classifier directly affects the performance of the system [169]. Inaccurate classification results will cause overall performance degradation. For example, in news summarization systems, if a non- news web page is classified as a news web page that has no connection with the news event that has to be summarized, it will be considered not relevant for summary generation, and thus not give good summarization result as it should be.

3.3 NEWS WEB PAGE CLASSIFICATION

Online newspaper websites have a standout amongst the most essential up to date information. Many websites provide day by day news in extremely different formats, and effective classification is required to get to and monitor this data in an automatic manner. In general news sites consist of thousands of web pages.

In general, a newspaper web site consists of thousands of web pages. The desired pages are the article pages. Thus, the first goal is to identify which pages are the article pages. The non-article pages, e.g. table of contents, advertisements, opinion or query submission forms, should be screened out.

News web page classification techniques use diverse information to classify a target web page: To classify the preferred news pages our approach identifies and explores common attributes that are commonly present in news websites. Maximum news websites are organized as (i) home page that shows some headlines of all sections. (ii) Numerous unit of pages that offer the headlines of diverse extents of interest like business, sports, entertainment, technology, politics etc. these different areas also contains some sub sections like national, international, market, cricket, football, science etc. (iii) pages that actually represent the news containing the title, author, related news link, date and body of the news. On the other hand web page URL is one of the most informative sources of information with respect to classification. URL of a web page is mainly content bearing, and it seems useful in making full usage of this resource. A classification method uses these attributes for the news web page recognition.

Therefore, we select URL attribute, structural attributes and content attribute for news web page classification which are least expensive to achieve and significant sources for classification. In this work, ten newspaper websites as sources of information are selected randomly. In this way, our approach for news web page classification depends on the collection of essential attributes and then uses a classification algorithm to correctly classify news pages disregarding the other pages or the non- news pages are screened out.

News web page classification has been extensively researched and several techniques have been applied with successful experimental results. In general, the research in this area can be classified according to the type of attributes. There are three main trends in attribute types: URL attribute, content attribute and structure attribute. URL is the more informatics sources of information for classification. Kan and Thi [106] present an approach to the segment the URL into meaningful chunks and ads components, sequential and orthographic features to model silent features. Many of their newly introduced features perform well on long URLs, typically found in an intranet setting. These features do not perform as well with typical web site entry points *i.e.*, just the domain name in contrast in our work we extract some more feature from URL like second level domain attributes, first level catalogue attributes that perform well on domain names also. A number of approaches have been proposed based on content attributes [100] [103] [167]. A set of words extracted from web documents are used as attributes for classification algorithm. In general keywords are extracted from the articles. The extracted keywords are used

to represent web documents and their classes. Typically, only some extracted keywords are selected as features or attributes for classification algorithms since the extracted word list is usually large and it is impractical for classification algorithms. Therefore some researchers use structural attributes of web documents for classification [97] [98] [102]. Cruz et al. [97] measured similarity between web pages based on the frequencies of HTML tags. Wong and Fu [98] generalized some knowledge from a hierarchical structure of web documents and used this knowledge to classify web pages. For web pages that describe the same type of information, it is common that they have similar structure. The structure can be used for identification of web pages that belong to the class.

Therefore in our work, we consider the limitation of all the previous research based on the individual attributes and used the all three attributes together to improve the classification accuracy.

3.3.1 ATTRIBUTES SELECTION

This work, for the automatic identification of news web pages depends on the selection of important attributes and then uses a classification algorithm to identify news web pages. Therefore we select the content, URL and structure attributes describe as:

3.3.1.1 Content Attribute

After selecting 750 html pages of news web pages and 275 HTML pages of non-news web pages, which was selected randomly from 10 different news websites, we observed that the occurrence of the “news” keyword in a webpage is an essential attribute for the news web page recognition. News in the news websites are classified as politics, sports, business, etc. in every category, there are also subcategories for example, the subcategories, that occur in the sports category are cricket, golf, tennis, football, hockey, etc. In business, market, share, economy, etc. We selected some keywords as content attributes of a news web page: News Center, article source, author, related news, interconnected subject, connected link and count how many times the term news appear in the HTML page, date.

3.3.1.2 URL Attributes

URLs are an extremely exquisite feature for learning. It is an important identification feature for web news; the URLs of news websites are often same structure. URL of news website contains both positive and negative attributes. For news web page identification positive attributes are more useful than negative attributes [170]. Second level domain attributes and first level catalog attributes come under positive attributes list.

3.3.1.2.1 *Positive Attributes*

- Second level domain attributes**:** Similar sections of different news web pages share related structure attributes. For example URLs of subsections of news web pages like business, tech and sports also have second level domain attributes such as "business", "tech", and "sports".
- First-level catalog attributes**:** URLs also contains first-level catalog attributes of news web pages such as "newspaper name" and news center. The First-level catalog attributes provides a vital basis for the recognition of the news web page.

3.3.1.2.2 *Negative Attributes*

- Bbs
- Blog
- Video
- Ads
- Campaign

3.3.1.3 Structure Attributes

News web pages encompass rich structure information that can increase the accuracy of a classifier if correctly used. Web pages of the same class would have similar ***structure*** which describes the format of information in the pages.

By analyzing the structure of different news web pages we observe that certain structure attributes contribute to news web page recognition, containing web page title and subtitle written as <title>, <Hn> tag and tag that form up a webpage's hierarchy. <title> tag of all news

websites are similar and contains web page title or news center and website information like newspaper name. tag contains the date and time feature of the webpage, which is necessary for news webpage recognition. The combined attributes of web news pages are shown in Table 3.1.

Table 3.1: Combined Attributes from Ten News Websites

Attributes		
URL attibutes	**Content attributes**	**Structure attributes**
Positive attributes Second level domain: news, tech, sports, country name, Business, economy, politics, science, Market, budget, gadgets, careers, world Games First- level catalog attributes: newspaper name **Negative attributes**: bbs, blog, video, ads, campaign,	News title Article source Author Top News Stories Latest News Related news Related link Related subject Sum up the number of times the term news appear in the HTML page Date and Time	Date and time feature in tag Has news center and newspaper name in <title> tag <a> tag contain top news or related news

3.4 EXPERIMENTAL DATASET

Experimental dataset for the classification of news web pages described in this paper depends upon the attributes from 10 different Indian news websites. These websites are named as Times of India, Hindustan times, NDTV, Indian Express, The Hindu, The Pioneer, India Today, Deccan Herald, The Asian Age, The Telegraph. To build up the several corpus section of news web pages are reviewed such as sports, politics, entertainment, technology, business. We select total 1025 web news pages posted from 10 July 2015 to 30 July 2015. For system evaluation, the dataset of 1025 news web pages are divided into two sets according to the attributes selection. We train Naïve Bayes algorithm chosen from WEKA (Waikato Environment for Knowledge Analysis) for classifying news web pages.

For the first set, dataset should be prepared by extracting the only two attributes as content and structure from five news websites name as "The Times of India, Hindustan times, NDTV, Indian Express and The Hindu" and then labeling the attributes with news or non-news labels. Then the labeled training data is used to train a learning algorithm. We reviewed total 556 pages in which 381 are news and 175 are non- news pages shows in Table 3.2.

Table 3.2: First Set of Dataset

Attribute selection from websites	News pages	Non-News pages
Times of India	80	36
Hindustan Times	70	40
NDTV	65	43
Indian Express	76	26
The Hindu	90	30

For the second set, dataset contains all three attributes Content, Structure and URL extracted from the all ten news websites name as "The Times of India, Hindustan times, NDTV, Indian Express and The Hindu, The Pioneer**,** India Today, Deccan Herald, The Asian Age, and The Telegraph". We select entire 1025 web pages in which 750 are news and 275 are non–news web pages shown in Table 3.3.

The model formed by a machine learning algorithm through training is tested on the test documents. Therefore dataset is divided into training and testing. For training purpose used five websites as, Times of India, Hindustan Times, NDTV, The Hindu and Indian Express. For testing purpose five different news websites name as The Pioneer, India today, Deccan Herald, The Asian Age and The Telegraph were used.

Table 3.3: Second Set of Dataset

Attribute selection from websites	News Pages	Non-News Pages
The Pioneer	85	25
India Today	82	20
Deccan Herald	87	15
The Asian Age	75	30

The Telegraph	40	10
Times of India	80	36
Hindustan Times	70	40
NDTV	65	43
Indian Express	76	26
The Hindu	90	30

3.5 LEARNING ALGORITHM

The choice of classifier depends on the practical requirements of the problem. We can usage any existing classification method to perform classification with these attributes. We classify using Naïve Bayes algorithm. This algorithm comes under probabilistic approach. For text classification applications and experiments generally Naïve Bayes classifier is used.

The naive Bayes classifier is much simpler and much more efficient than other supervised learning. This is largely because of the independence assumption which allows the parameters for each feature to be learned separately. In the case of document classification, number of features is document classification is usually proportional to the vocabulary size of the training document set. This number can be quite large in many cases so the efficiency of the naive Bayes classifier is a major advantage over other classification techniques [171]. The basic idea is to use Naïve Bayes classifier is the joint probability of words and categories to assess the probabilities of the classifications given a document [172].

3.6 PROPOSED APPROACH

To classify news web pages correctly, Naïve Bayes classifier is used to recognize news pages from non-news pages. The overall process of news web page classification has been shown in Figure 3.1. The steps of the approach are as follows:

Step 1: Select online news websites randomly.

Step 2: Generate a dataset containing content, URL and structural attributes.

Step 3: Use Naïve Bayes classifier to identify the news pages from non-news pages.

Step 4: Evaluate the performance of Naïve Bayes classifier in terms of precision using WEKA tool.

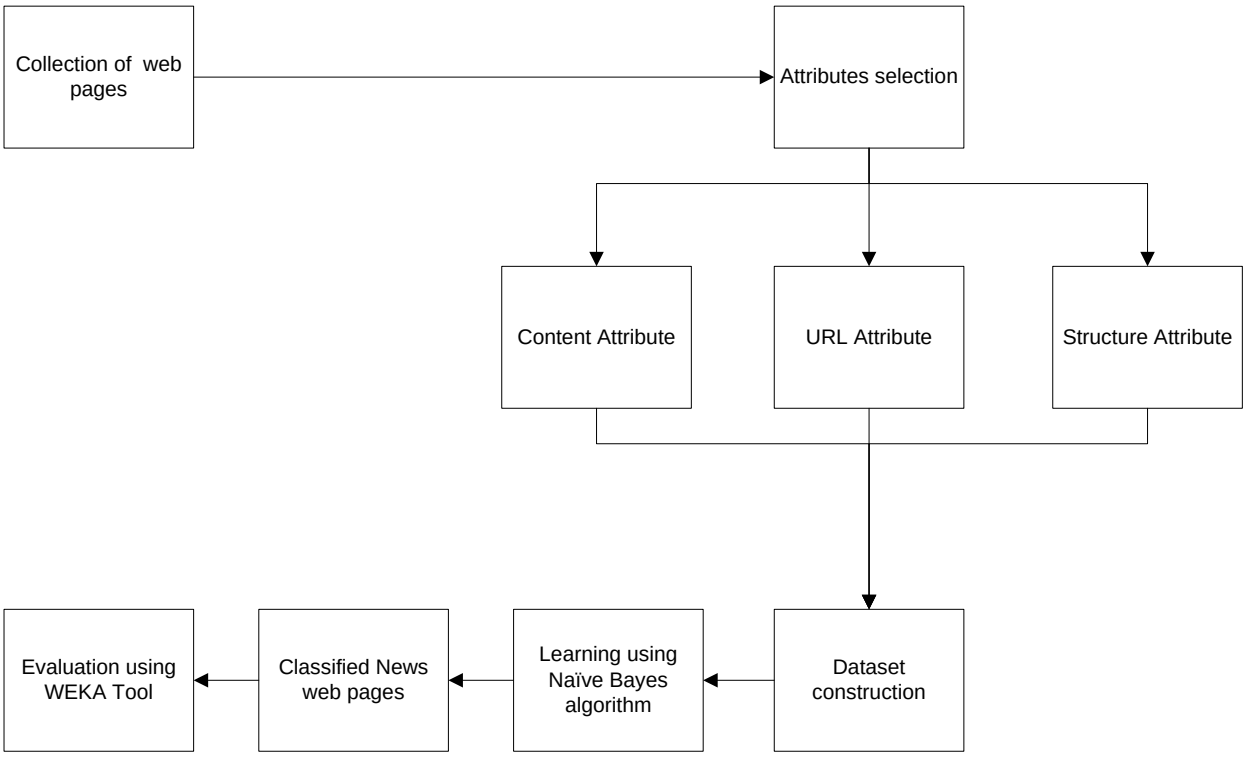


Figure 3.1. Workflow of News Web Page Classification

3.7 EXPERIMENT AND RESULTS

In our experiment the performance of Naïve Bayes classifier is evaluated using precision analysis. Precision is defined as the fraction of the number of correctly classified documents over number of documents which classifiers classified to this category so the precision is defined as in equation (3.1).

$$P = \frac{a}{a+b} \tag{3.1}$$

Where a = Number of correctly classified documents.

And b = Number of incorrectly classified documents.

For this purpose WEKA has been used. It is a traditional machine learning tool [173]. Based on the attributes from different websites, we conducted two set of experiments accordingly.

As we discussed in earlier section 3.4, dataset is divided into two sets. In the first set of experiments, randomly only two attributes content and structure are extracted from five news websites, and training and testing accomplished with the dataset from the same websites. The labeled collection of pages is divided into two parts, namely 70% are considered as training data and 30% are testing data. The training process requires that a set of categories has been defined, and the training data need to be somehow labeled with their respective category. The Naïve Bayes classifier must be trained on a training set of web pages, and preferably evaluated on a smaller testing set. The evaluation is useful for determining the optimal values of any parameters of the classifier on the available training data. The experimental results are shown in Table 3.4.

In the second set of experiment, extracted all the three attributes (content, structure and URL) from ten news websites, dataset is further divided into training set and testing set. From the ten news websites, randomly five websites are selected for training purpose and five different websites for testing (as discussed in section 3.4) to show the variation in the results. Evaluation of outcomes using all three attributes where different websites are used for training and testing are shown in Table 3.5.

Table 3.4: Precision Values using First Set of Dataset

Feature selection from websites	Total Pages	Training	Testing	Correctly Classified	Incorrectly Classified	Precision
Times of India	116	81	35	34	1	0.971
Hindustan Times	110	77	33	31	2	0.939
NDTV	108	76	32	31	1	0.968
The Hindu	120	84	36	34	2	0.944
Indian Express	102	71	31	30	1	0.967

Table 3.5: Precision Values using Second Set of Dataset

Feature selection from websites	Total Pages	Training	Testing	Correctly Classified	Incorrectly Classified	Precision
Times of India and The Pioneer	226	116 (Times of India)	110 (The Pioneer)	99	1	0.990
Hindustan Times and India Today	212	110 (Hindusta n Times)	102 (India Today)	96	2	0.979
NDTV and Deccan Herald	210	108 (NDTV)	102 (Deccan Herald)	94	4	0.989
The Hindu and The Asian Age	225	120 (The Hindu)	105 (The Asian Age)	91	6	0.958
Indian Express and The Telegraph	152	102 (Indian Express)	50 (The Telegraph)	48	2	0.980

Results from the both set of experiments show that precision value using second set of dataset where all the three attributes are extracted from the ten different news websites and five websites are used for training and five different websites are used for testing; are better than the first set of dataset where only two attributes are extracted from the five websites and same five websites are used for training and testing purposes. From these experimental results we can say that combination of all the three attributes "content, structure and URL" gives better results rather than the two attributes alone.

3.8 EXPERIMENTAL ANALYSIS

Extraction of the main attributes from news webpages is important for correct classification of web pages. According to An Chy et al. [109] classification of news web pages based on their content shows good results in correct classification. Patil & Pawar [174] also focused on content based classification. In their work, they used the Naïve Bayes classifier for the home page of a website for classification into industry type category. Their results show the precision value of 89.09% and recall 89.04%. Kan and Thi [106], shows that URL based classification are very effective both in space and time. They used the standard dataset TREC, and WebKb to perform experiments. Tongchim et al. [108] used only the structural features of web documents and applied machine learning method for the identification of news web pages

from the non-news web pages. They used Thai web corpus for their experiments; their results show the correct classification of documents. Therefore, attributes must be selected carefully for recognizing news web pages. We combine the three attributes and find better results. According to Wu et al. [170], any existing method can be used to perform classification so we used Naïve Bayes algorithm in this work. It gives better results for our dataset.

The classification results using the Naïve Bayes algorithm showed in Table 3.4 and Table 3.5. From Table 3.4, it can be observed that the precision value using Naïve Bayes algorithms for the first set of datasets, where only two attributes has been selected from the five websites and training and testing conducted with the same websites (Times of India, Hindustan Times, NDTV, The Hindu and Indian Express) as 0.971, 0.939, 0.968, 0.944 and 0.967 respectively.

Table 3.5 shows the precision value using Naïve Bayes algorithm for the second set of dataset, where all three attributes has been selected and attributes of five websites (Times of India, Hindustan Times, NDTV, The Hindu and Indian Express) are used for training and five different websites (The Pioneer, India today, Deccan Herald, The Asian Age and The Telegraph) are used for testing are 0.990, 0.979, 0.989, 0.958 and 0.980 respectively.

Previous works on news web page classification show that [100] [103] SVM and decision tree are also used for the classification. In WEKA for SVM, SMO is used and J48 for the decision tree [175]. Therefore we have also train recognition classifiers with SMO and J48 on the same features. The experimental results demonstrate that the three classifiers built by Naïve Bayes, SMO and J48 all have a high precision for web news classification, while Naïve Bayes classifier shows high precision value for both set of experiments.

Comparison results of Naïve Bayes, SMO and J48 algorithms using the first set of dataset in term of precision shown in Table 3.6. According to it, precision of Naïve Bayes classifier for the first set of dataset is 0.951 which is better than SMO and J48 as 0.859 and 0.766 respectively.

On the other hand, evaluation of comparison results of Naïve Bayes, SMO and J48 algorithms using dataset of second set, containing different websites for training and testing regarding precision are shown in Table 3.7. According to the table, Naïve Bayes also shows the

high precision value for the second set as 0.965, which is better than SMO and J48 as 0.847 and 0.817 respectively.

Table 3.6: Comparison Among Three Algorithms using the First Set of Dataset

Algorithms	Times of India	Hindustan Times	NDTV	The Hindu	Indian Express	Average Precision
Naïve Bayes	0.971	0.909	0.968	0.944	0.967	0.951
SMO	0.763	0.882	0.957	0.896	0.799	0.859
J48	0.946	0.856	0.781	0.819	0.432	0.766

Figure 3.2 shows the comparative analysis of these three algorithms using first set of dataset by graphical representation.

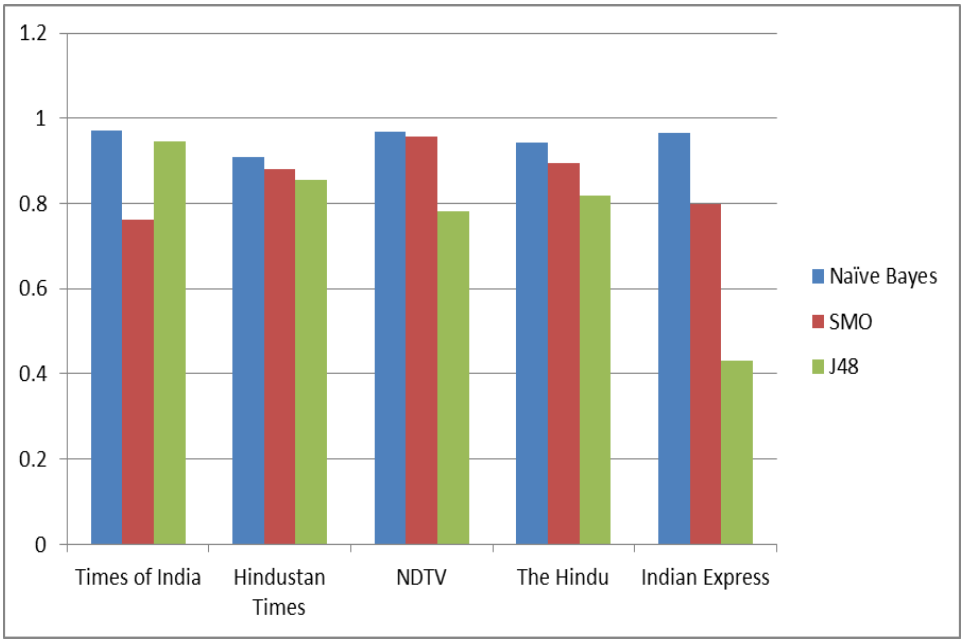


Figure 3.2. Precision Comparison of Three Algorithms using First Set of Dataset

Graphical representation of precision comparison of all three algorithms using second set of dataset is shown in Figure 3.3.

Table 3.7: Comparison Among Three Algorithms using the Second Set of Dataset

Algorithms	Times of India and The Pioneer	Hindustan Times and India Today	NDTV and Deccan Herald	The Hindu and The Asian age	Indian Express and The Telegraph	Average Precision
Naïve Bayes	0.990	0.979	0.959	0.938	0.960	0.965
SMO	0.738	0.817	0.930	0.813	0.939	0.847
J48	0.869	0.601	0.882	0.840	0.896	0.817

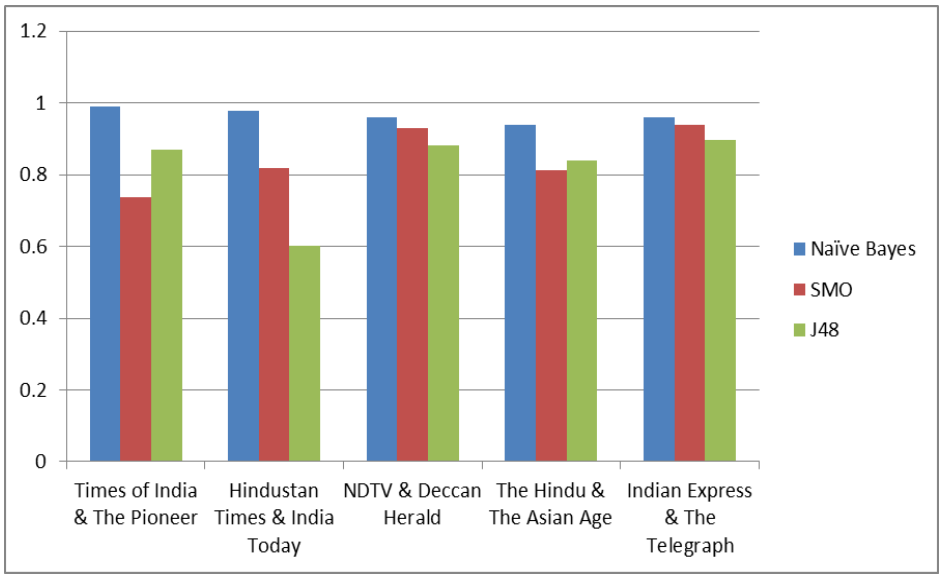


Figure 3.3. Precision Comparison of Three Algorithms using Second Set of Dataset

The experimental results show that from an extensive usage of news web page content, URL and structural attributes in the attributes selection Naïve Bayes classifier gives accurate results when identifying news web pages. Date attributes second level domain and first level catalog are the most predictive attributes. In addition, the frequency of the term news occurs on HTML page is significant for content attributes for news webpage.

3.9 SUMMARY

In present work, we have proposed news web page classification approach and identified important attributes. The combination of three attributes as content, URL and structure attributes are used for the training and testing purposes and Naïve Bayes classification algorithm classify the news web pages from non-news web pages. We conduct two set of experiments, in the first set of experiment only two attributes are chosen from the five news websites and training and testing conducted on the same websites while in second set of experiment, all three attributes are selected from the all ten news websites and training dataset and testing dataset are chosen from the different websites. We also compare the Naïve Bayes algorithm with SMO, and J48 algorithms for both set of dataset, results show that Naïve Bayes algorithm performs better than the other two algorithms for the dataset. Experimental results show that combination of the three attributes can yield more promising results than individual attributes.

Chapter 4

Content Extraction From News Web Pages

CHAPTER 4

CONTENT EXTRACTION FROM NEWS WEB PAGES

4.1 INTRODUCTION

In this chapter, we propose an approach for extracting the main content from news web pages. Our approach is based on the concept of tokenization of HTML page, these tokens construct the tag tree; web pages from different websites are parsed into Tag Tree and generated a template from each web pages and discover matching patterns and multiple sequence alignment. It finds and removed shared token sequences from the web pages until the relevant information is extracted from them.

4.2 CONTENT EXTRACTION

There is a vast amount of information available on the web, but most of the information is not in a form that can be easily used by the user. Efficient access to the relevant information within the huge amount of information needs major efforts. The process of content extraction is defined as the extraction of relevant content from massive data, such as text, database, semi-structured and multimedia documents. Efficiently extracting high quality content from news web pages is a challenging yet important problem in the field of information retrieval and news summarization.

Web pages have their own underlying embedded structure in the HTML language [122]. A major problem in news content extraction is mining useful information from web because news web pages not only contains the actual news content but also some noisy content like advertisement, comments and branding banners etc. Therefore, by analyzing several news websites, the actual news content is just half, and noisy content occupies nearly half of the page. If a content extraction method is directly applied to these pages, it is possible to lose focus on the main topics and important content.

News articles are unstructured documents, whose relevant information is pieces of free text. To extract the relevant news from the whole web page, our approach identifies and searches common characteristics that are usually present in the news web pages. Like most news websites have some common structure namely (i) a home page that presents the important headlines from all fields. (ii) Several section pages divided in different areas of interest like business, sports, National, International, Technology etc that provide the related headlines, and (iii) Pages that actually present the news, containing the title, author, date and body of the news. Our approach is based on the basic assumption that the news web pages content is divided into tokens where tokens represent the HTML tags like <head><title><script> etc.

In this work, we extracted the core content from a number of news web pages and these news web pages come from ten different news websites. We mainly deal with news pages written in English.

Identification of actual news content from news web pages is relatively easy task for the human being, who can identify just by visual inspection; however it is hard problem for machines. Our approach not only extract the relevant text passage from the given news website but also the fetching of the entire website content, and the extraction of the relevant content.
Content extraction that depend on extraction rules do not usually adapt well to changes to the web. When the set of extraction rules is handcrafted or learnt, the web keeps growing and it is uncommon that changes may invalidate the existing extraction rules. Therefore some authors work on semi-automatic extraction rules. Our work does not rely on extraction rules like previous approaches [116] [117] discussed in chapter 2. It requires input web pages and translated into Tag Tree. It works on two or more web documents and compares them to obtained shared patterns that are likely to provide relevant information. The idea of identifying shared pattern relies on tree matching and determines which are equivalent to one another and apply filtering algorithm to filter out irrelevant content. We have conducted experiments with 500 news web pages from ten different news websites and our results discussed in section 4.5 confirms that our approach can achieve good precision and recall values for extract meaningful content from news web pages.

4.3 PROPOSED ALGORITHM

The idea of our algorithm originated from the earlier work of Sleiman & Corchuelo [123] in 2013. In their work, TextSet has been used for the web information extraction whereas in this work we use ContentSet for the extraction of news web page content using Tag Tree. Our algorithm is divided into four components. First one is Tag Tree [121] that is used to measure the similarity between the templates of web pages. The second component is an extractor that returns a list of ContentSet that contains as much potential information as possible. The third component is Pattern matching that finds out repetitive pattern and the fourth component is a filter that filters out undesired patterns and return candidate pattern. Our proposed approach works on a collection of web documents, which we denote as ContentSet (CS). A content set is a set of contents that are sequences of HTML tags. The implementation and experiments were based on these HTML tags. Tag Tree is constructed to describe a web page, where tree nodes are defined by HTML tag and text. HTML tags are the basic components for document presentation and convey certain structure information.

Referring to Figure 4.1, a flowchart of the content extraction process is shown. When a user submits an html page, the page is tokenized, which is responsible for segmenting the input page into simple tokens that represent either script blocks, style blocks or html tags by using Python NLTK word tokenization [176]. It provides a number of tokenizers in the tokenize module. The Tag Tree is constructed by these tokens. The pattern matching then uses the Tag Tree to discover repetitive patterns. The repetitive patterns are forwarded to filtering, which filter out undesired patterns and finally the extracted content is found.

The process of content extraction for our work utilizing different steps can be combined and presented in the form of following algorithm.

Algorithm 1

1. TT = Tag Tree (CS, html tags)
2. extract (CS: ContentSet; End, Start, Max, Min)
3. l = extract (CS, TT, End, Start, Max, Min)
4. m = PatternMatching (l)

5. result = filter (m)
6. return result

The algorithm works in four steps: at line 1, we invoke Tag Tree. At line 2, we invoke algorithm extract, which makes an attempt to extract the information that varies from document to document. At line 4, we invoke the pattern matching algorithm. This algorithm searches for the shared patterns of size End, End-1,….Start, where Start is the first document and End is the last document in the ContentSet. If Start >1 or End is less than the size of the input document, then the search has a preference that may lead to situation in which pattern matching algorithm return information that actually belong to the template, therefore at line 5, we invoke filtering algorithm.

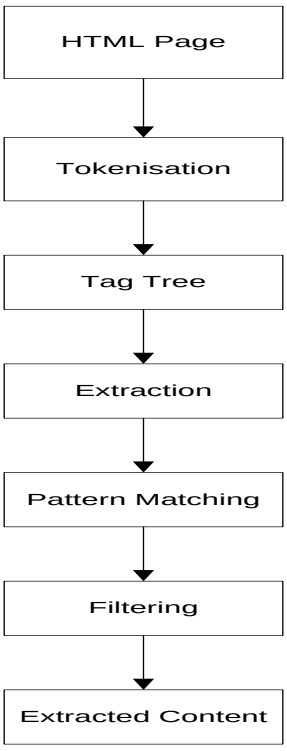


Figure 4.1. Process of Content Extraction

4.3.1 TAG TREE

We used the concept of Tag Tree [121] in our work. HTML in a web page is parsed as a Tag Tree in our work. Tag Tree is a hierarchical implementation of tags and uses DOM (Document Object Model) to keep developed. One tag leads to page that have a 'child tag' which further leads to 'sub child tag'. Tags are denoted as nodes in tag tree. Attributes in Tag Tree are also regarded as nodes.

Following rules are used for the construction of Tag Tree.

1. There are mainly three type of node in the Tag tree, which is summary node, text node and Tag node.
2. Entire content of news with its subtrees are wrapped in a pair of node such as <HTML></HTML> is summary node.
3. Tags and text between a pair of tags, such as <body></body> are all children nodes of the Tag tree.
4. All the content between a pair of <script> tags is a text node as the only child of <script> node.
5. All attributes of a node which are parsed in order, will be inserted into the attribute map.
6. A tag ended with "/>" is a node with self-closing flag is true such as <frame.src= "sun.htm"> of a XHTML.

Figure 4.2 shows the source page of news web page. We create Tag Tree with the help of this source page.

```
<html>
<head>
    <meta charset="utf-8" />
    <meta name="viewport" content="width=device-width, initial-scale=1.0, user-scalable=yes" />
    <meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
    <meta http-equiv="Content-Language" content="en">
    <meta name="copyright" content="2015 Kashmir Images. All rights reserved.">
    <meta name="distribution" content="global">
    <meta name="robots" content="index, follow, noarchive">
    <meta property="twitter:site" content="kashmirimages">
    <meta property="twitter:card" content="summary">
    <meta property="twitter:domain" content="http://www.dailykashmirimages.com">
    <title>Four killed as militants ambush army convoy in Shopian</title>

        <meta property="og:title" content="Four killed as militants ambush army convoy in Shopian">
        <meta property="og:description" content="
        Srinagar, Feb 23 (PTI) Militants ambushed an army convoy in Shopian district of Kashmir today, killing three soldiers and a woman.

        The militants attacked the security forces ">
        <meta property="og:url" content="http://www.dailykashmirimages.com/Details/132340/four-killed-as-militants-ambush-army-convoy-in-shopian">
        <meta property="og:image" content="http://www.dailykashmirimages.com/upload_images/upload_images_articles/cb100e77-f0d4-4e26-b1b3-92a1fe5804f6.JPG">
        <meta property="twitter:title" content="Four killed as militants ambush army convoy in Shopian">
        <meta property="twitter:description" content="
        Srinagar, Feb 23 (PTI) Militants ambushed an army convoy in Shopian district of Kashmir today, killing three soldiers and a woman.

        The militants attacked the security forces ">
        <meta property="twitter:url" content="http://www.dailykashmirimages.com/Details/132340/four-killed-as-militants-ambush-army-convoy-in-shopian">
        <meta property="twitter:image" content="http://www.dailykashmirimages.com/upload_images/upload_images_articles/cb100e77-f0d4-4e26-b1b3-92a1fe5804f6.JPG">

    <link href="http://fonts.googleapis.com/css?family=Open+Sans:400,600,700" rel="stylesheet" type="text/css" />
    <link href="/Content/bootstrap.min.css?desktop" rel="stylesheet" />
    <link href="/Content/jquery.bxslider.css" rel="stylesheet"/>
<link href="/Content/perfect-scrollbar.css" rel="stylesheet"/>
<link href="/Content/side_menu.css" rel="stylesheet"/>
<link href="/Content/smoothDivScroll.css" rel="stylesheet"/>
<link href="/Content/style.css" rel="stylesheet"/>
<link href="/Content/style_business.css" rel="stylesheet"/>
```

Figure 4.2. Source Page

Figure 4.2 shows the structure of a web page. By manual inspection we found that all news web pages show the similar structure and contains following tags <html>< head><meta><title></title><link ref><script></script> <p></head></html>. With the help of these tags we construct the Tag Tree.

Figure 4.3 shows the Tag Tree of a news web page. In this tree <HTML> is a summary node. <body> represent the children node are the Tag node of the tree. Tags come under the <script> tag denoted in text node.

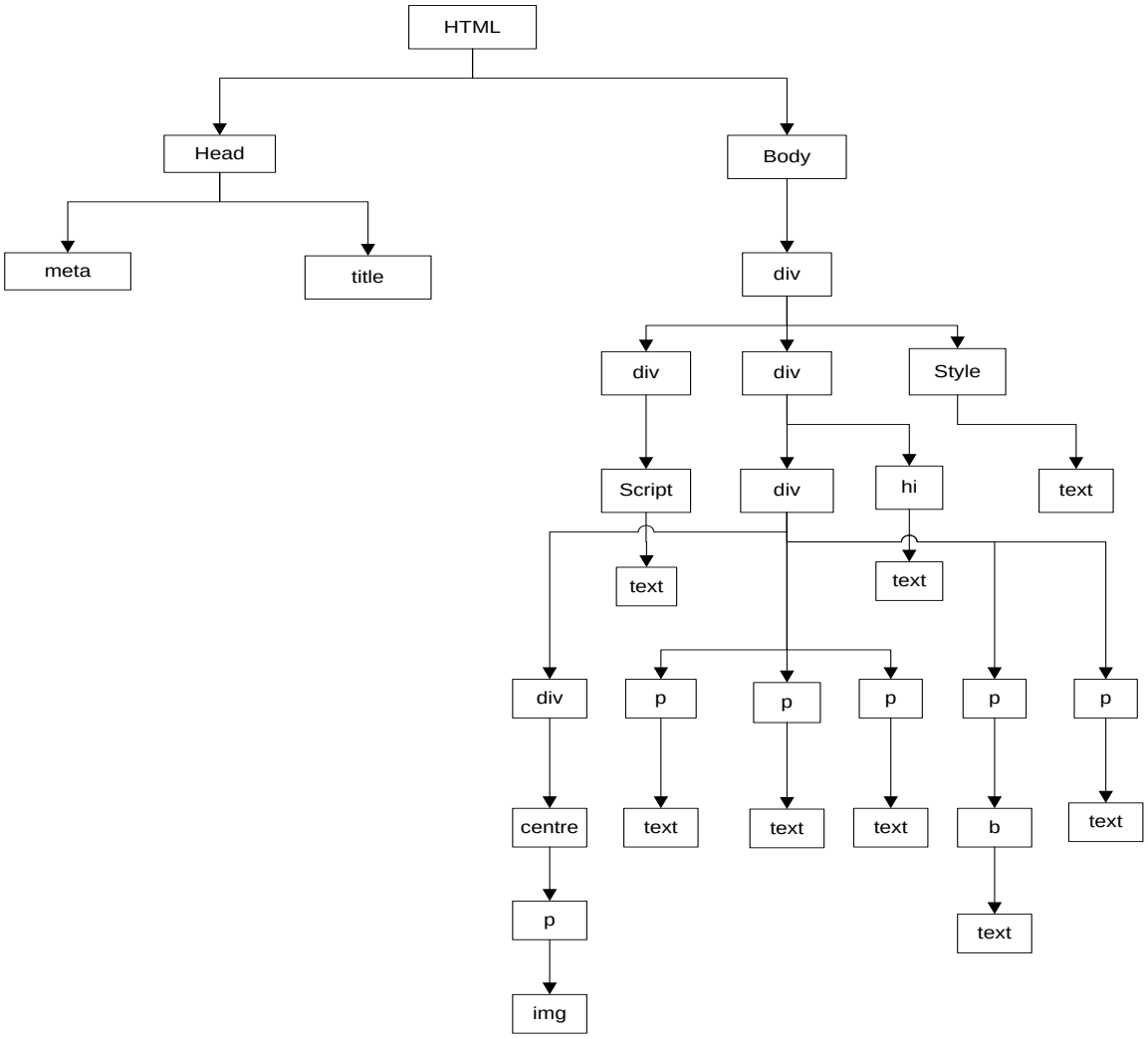


Figure 4.3. Tag Tree

4.3.2 EXTRACT ALGORITHM

Extract algorithm searches for shared patterns of size Start down to End in a Contentset. We find the shared patterns using Tag Tree. Algorithm 2 presents the algorithm extract that lies on the Contents, which makes an attempt to extract the information that varies from document to document. It works on a collection of web documents which we denote as ContentSet (CS). Intuitively, a Content set is a set of Contents which are sequences of tokens. Content is not bound with a particular tokenization schema. Our implementation and our experiment were carried out using a simple tokenization schema according to which tokens represents Tag set and we construct a Tag Tree as shown in Figure 4.3. The algorithm extract can search the tag tree to

find all occurrences of the extraction pattern. It works on a ContentSet CS where Start is the first document and End is the last document in the ContentSet, and Min and Max are the minimum and a maximum pattern size is which returns a list of ContentSet that should contain eventual information. The main loop 3 to 15 iterates over the entire documents in the ContentSet from start down to end. The inner loop at lines 4 to 14 searches for a shared pattern of that size. In Figure 4.3 the longest shared pattern we used is of size 7 tokens <html><head><title>News title</title></head></body></html>. In this algorithm variable buffer act as a queue in which we initially put the CntentSet on which the algorithm has to work, and at line 7 ContentSet is removed from the buffer. Algorithm searches for shared pattern of a given size in the ContentSet. If shared pattern is found then those patterns are added to result. If no shared pattern is found, then the original ContentSet is added to the buffer. Once the inner loop finishes, the result contains all the new ContentSet that has been produced, and it is transferred to the buffer variable so that the algorithm can search for new shared patterns of a smaller size, if possible.

Algorithm 2

```
1.  extract (CS : ContentSet ; Start, End,  Max, Min, TT: Tag Tree)
2.       buffer = <CS>
3.      for each = Start down to End do
4.          for size = Max down to Min
5.           result = <>
6.          while buffer ≠<> do
7.              CS = dequeue (buffer)
8.           if TT= SharedPatterns(CS, size) then
9.                enqueue (result, CS)
10.           else
11.               enqueue (buffer, TT)
12.          end
13.      end
14.    buffer = result
15.  end
16. return result
```

This algorithm of extract news is to extract entire content from news web page. The input of it is a news web page. The extracted document is shown in Figure 4.4.

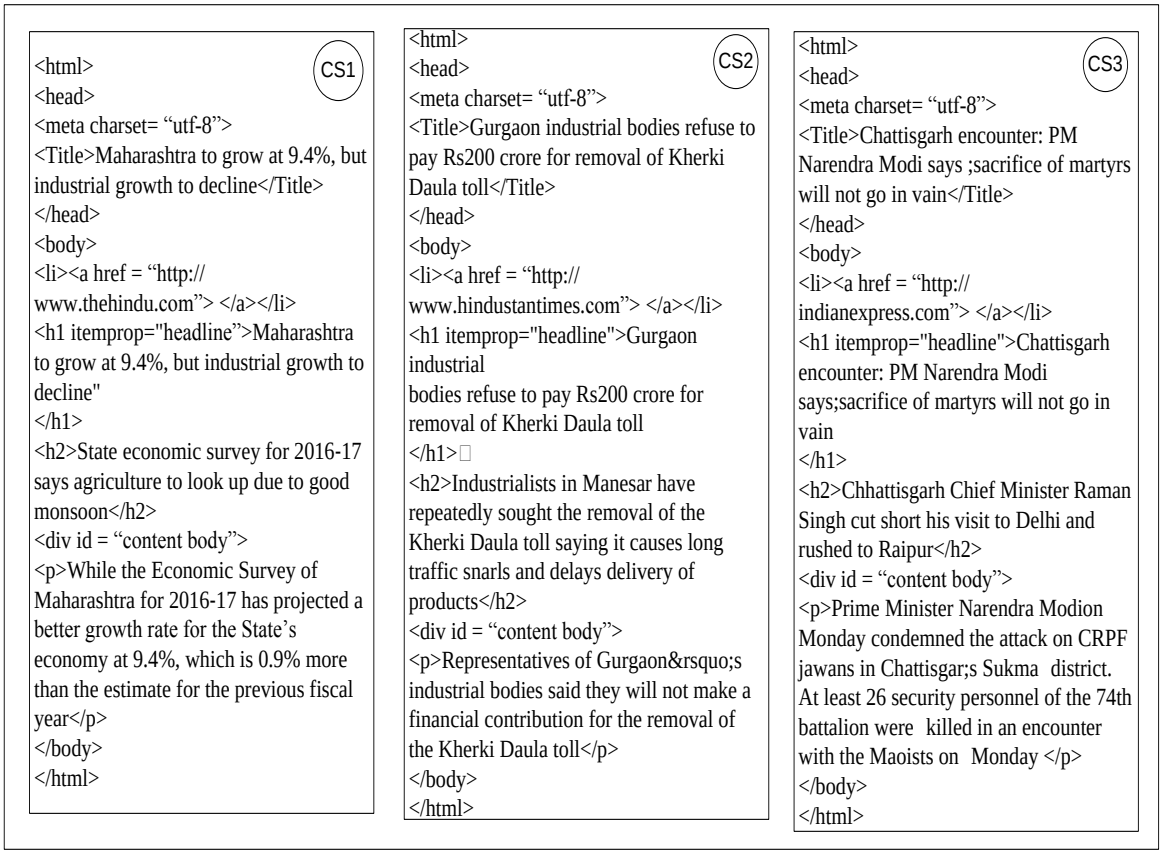


Figure 4.4. Extract Result

4.3.3 PATTERN MATCHING

After the extraction, user may select target pattern that contain desired information. The motivation of our algorithm is based on the observation that news web pages placed the desired information in a structure having a particular alignment and forms similar patterns. All the leaves in a Tag Tree share common Tags. In the Tag Tree each path from root to the root of the subtree represents a similar sequence or pattern in the input. Therefore, to find similar patterns, we need to examine the path to determine whether they are maximal similar or not. Algorithm 3 presents the pattern matching algorithm. It works on a ContentSet CS, a path of size s, which is supposed to be the shortest non-empty path in Tag Tree. The goal of this algorithm is to find and match similar patterns according to the path of the Tag Tree that occurs in every content in CS. In this algorithm, at lines 3-11, i iterates from 0 until size s. as long as no matching pattern is found. The

actual search is performed in the inner loop at line 6-10; in this loop the algorithm iterates over every content in the input ContentSet and finds all the matching patterns that start at position i and has size s. The algorithm returns a list of matching patterns in the ContentSet.

Algorithm 3

```
1.  SimilarPatternMatching (CS: ContentSet ; path: Content; TT: Tag Tree)
2.      Found= false
3.      For i = 0 until size (path) – s while not found do
4.          Result = {}
5.          found= true
6.          foreach Content in CS  while found do
7.              If TT  = findsimilarPattern( Content, path, s) then
8.                 found = size (matches) > 0
9.              else
10.                found = size (no matches)
11.                result=  found (TT)
12.         end
13.     end
14. return result
```

In Figure 4.5 to search for a pattern of size 4; we supposed that base is the shortest content in CS1, CS2 and CS3. The algorithm first searches for site link
<ahref= “site link”>
in every content in CS1, CS2 and CS3 and found it. Then it searches for <h1= “headline”></h1>, <h2>sub headline</h2> and News content which is found in every ContentSet CS1, CS2 and CS3. As a conclusion <a>, <h1>, <h2>, are the matching pattern in the content set. The Tag Tree of pattern matching is shown in Figure 4.6. All the leaves in a Tag Tree share a common prefix, all the three news web pages shows the common Tag Tree. Leaves represent the repeated sequence of input.

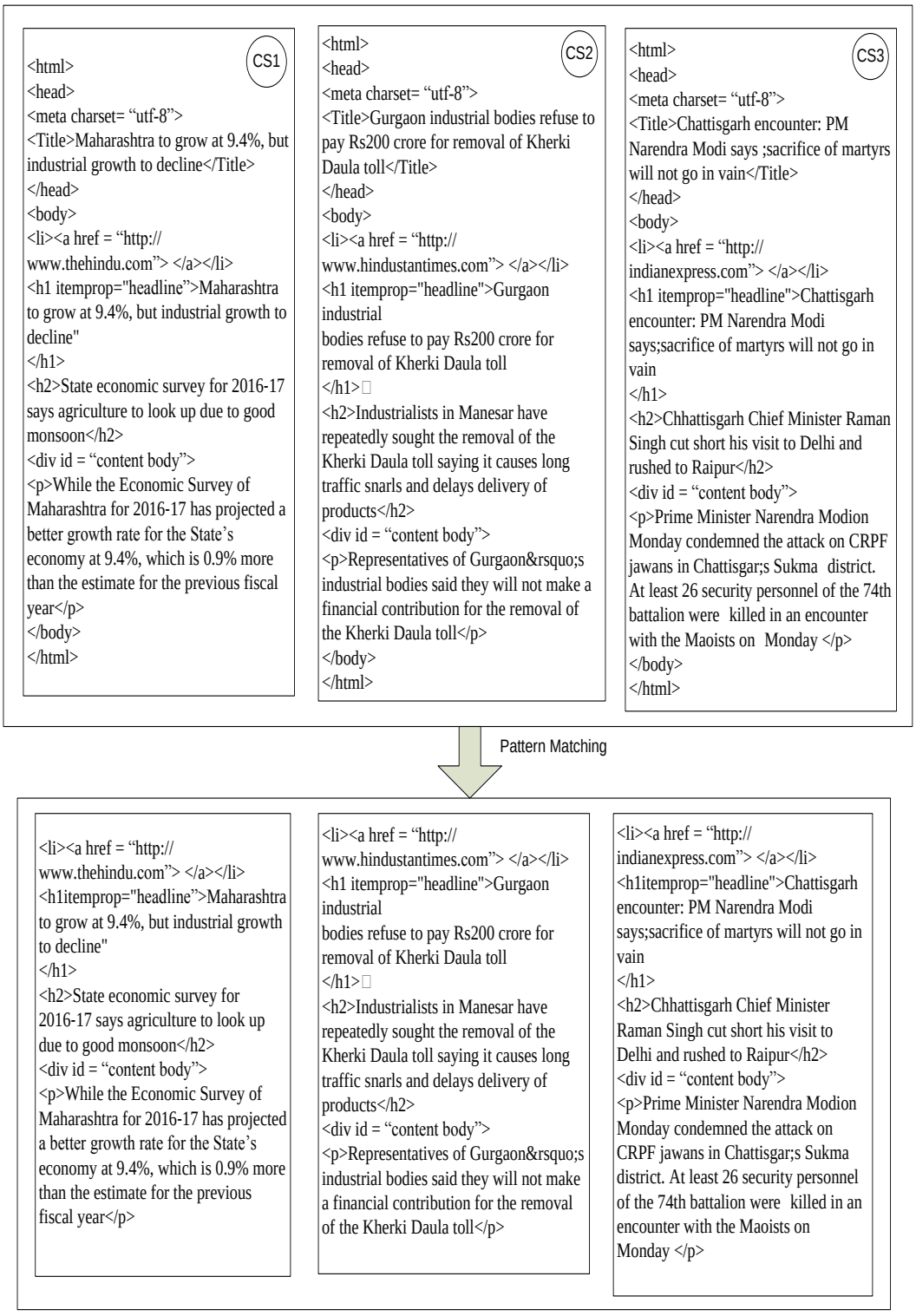


Figure 4.5. Searching for Pattern Matching Using <a href=”site link”> as a Base.

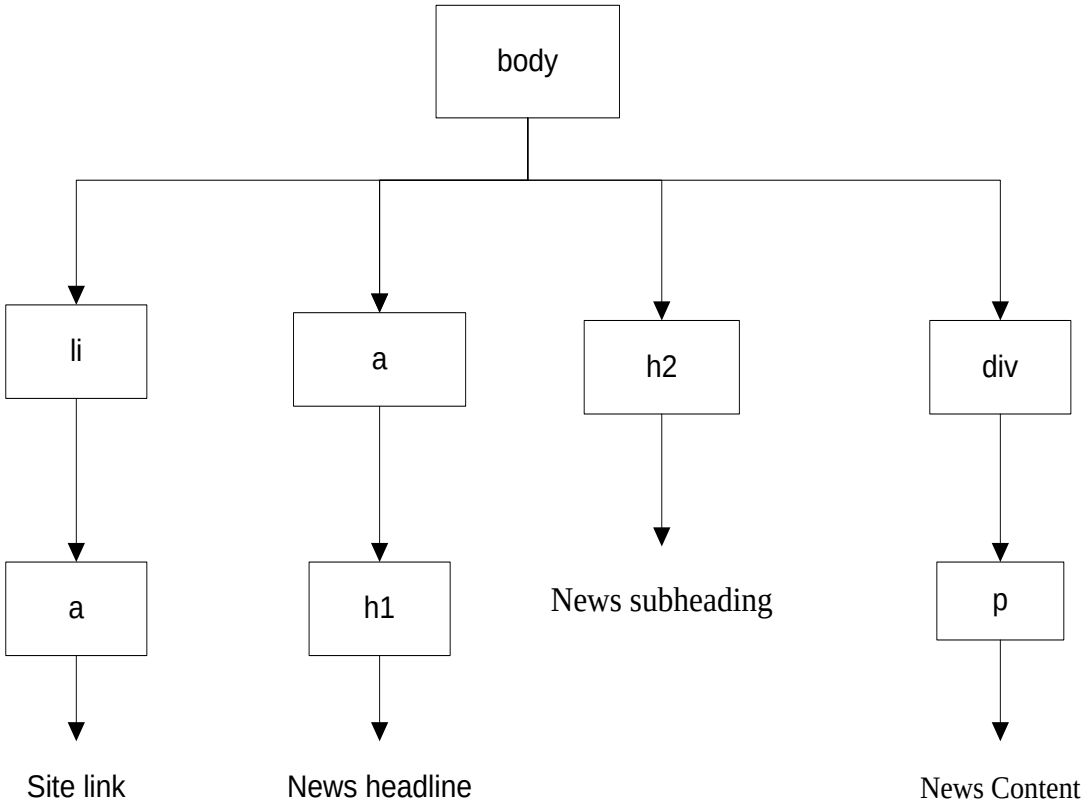


Figure 4.6. Tag Tree of Pattern Matching

4.3.4 FILTERING

After extraction and pattern matching filtering algorithm is applied. Typically a news web page usually contains a large number of similar patterns, not all of which contain useful information. To filter out undesired similar patterns, filter algorithm is used. For the filtering of undesired patterns we use two criteria, compactness and variability. Compactness is a measure of the density of maximal similarity. It can be used to filter out those patterns which are far apart beyond a given bound. Density is defined by Chang and Lui [177] as

$$(k-1)\times\frac{|\alpha|}{T_k - T_1} \tag{4.1}$$

Where $|\alpha|$ is the length of α in number of tokens. $T_1 \ldots T_k$ are the token sequences. We set the density value to 0.8, therefore the similarity value greater than the given value be qualified for extraction.

Variability is another criterion which filters out the pattern which shows no variability in the patterns. We denote the size of the pattern P as n, and then this loop iterates n times. In each iteration, algorithm checks the variability of the current ContentSet Tag tree to every other in order to determine whether the ContentSet has variability or not.

Algorithm 4 presents the filter algorithm. Extraction algorithm returns a list of ContentSet that are supposed to contain compactness (density of maximal similarity) and the variable information in the initial ContentSet. In this algorithm, the main loop at lines 3-7 iterates over the list of input ContentSet and simply removes those in which has compactness value is less than the default value (0.8) and shows no variability in the patterns. Let CS be the ContentSet, and d is the compactness (maximal similarity density) of ContentSet. The result of filter algorithm is shown in Figure 4.7.

Algorithm 4

1. Filter (CS: ContentSet, TT: Tag tree, P: Pattern size; d: compactness (Maximal similarity density))
2. Result = ⟨⟩
3. Foreach Tag Tree in CS do
4. If d >= 0.8 then
5. add d to result
6. If CS has variability then
7. add CS to result
8. end
9. end
10. return result

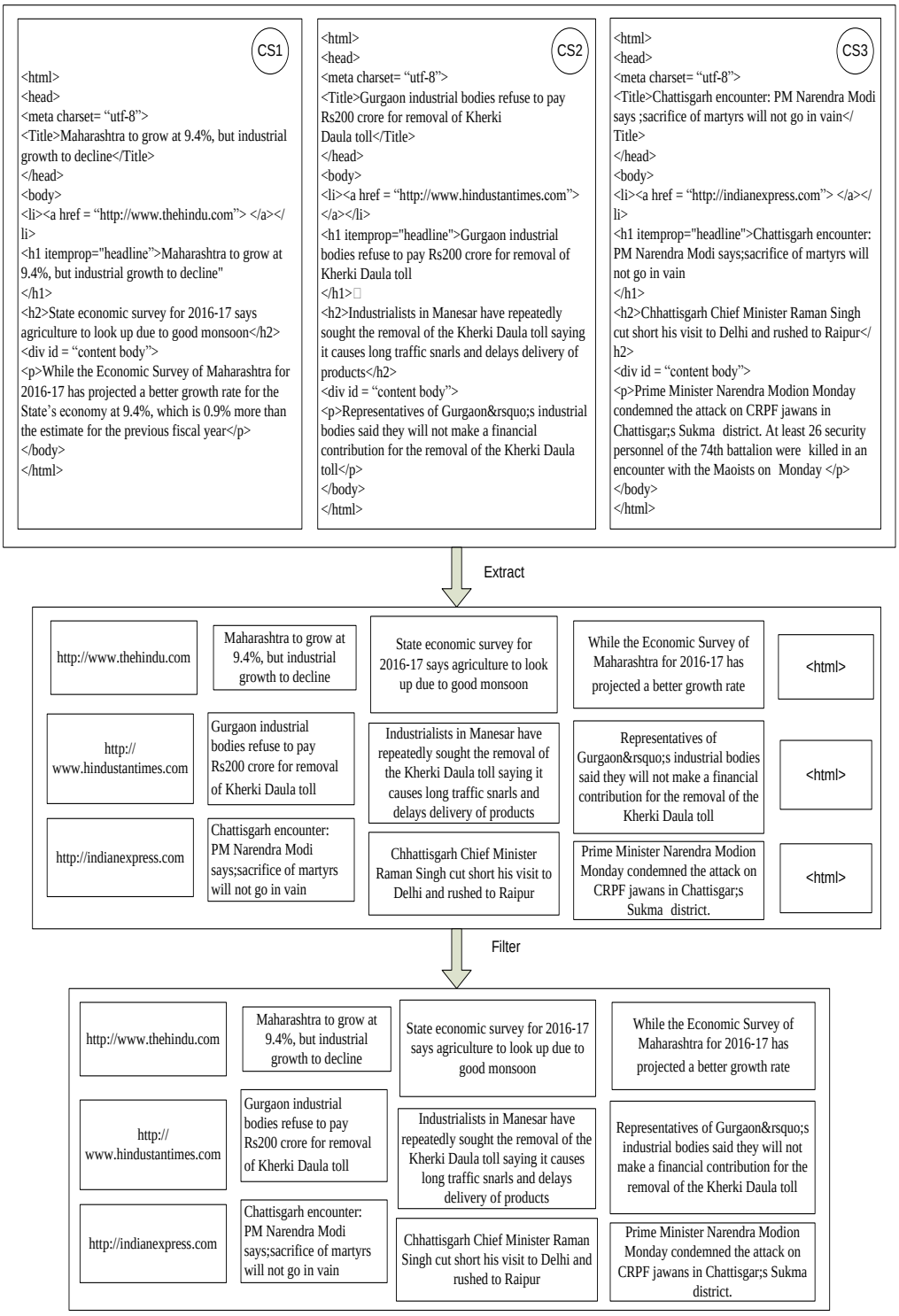


Figure 4.7. Filtering Result after Extraction

4.4 EXPERIMENTAL DATASET

The dataset named ContentSet (CS) used in our experiment contains a total of 500 news web documents. For experiments, we collected web news pages from 10 news websites, 50 articles from each website written in English. The sites are shown in Table 4.1. News web pages belongs to different categories like market, business, India, Tech, Nation, Science & Environment, politics, world, Entertainment, Sports. Each category was randomly selected from Google search engine between December 2016 to March 2017. We downloaded 50 web pages from each web site.

Table 4.1: News Websites

S. No.	News website	URL	Category	Number of Documents
1	The Economic Times	http://economictimes.indiatimes.com/	Market	50
2	The Hindu	http://www.thehindu.com/	Business	50
3	The Times of India	http://timesofindia.indiatimes.com/	Sports	50
4	NDTV	http://www.ndtv.com/	India	50
5	Hindustan Times	http://www.hindustantimes.com/	Tech	50
6	Indian Express	http://www.indianexpress.com/	Nation	50
7	News18	http://www.news18.com/	Politics	50
8	The Pioneer	http://www.dailypioneer.com/	World	50
9	Deccan Harald	http://www.deccanherald.com/	Entertainment	50
10	The Asian Age	http://www.asianage.com/	Science& Environment	50

News web documents preprocess by fixing their HTML code by HTML tidy [178]. It fixes web documents doctype declarations, adds missing end tags, and reports on unknown attributes if essential. Our dataset were gathered from real world news websites, they usually contained errors in their HTML code. Table 4.2 shows the results we have gathered regarding a subset of common HTML errors that are reported by HTML tidy. The full report is too large to reproduce in the work. Our only purpose to use HTML tidy is to make it clear that we have dealt with actual documents.

Table 4.2: Common Errors

S. No.	Error
1	Error: <!DOCTYPE> is missing
2	Error: <s6> is not recognized
3	Error: <j> is not recognized
4	Error: <bw> is not recognized
5	Error: <zs> is not recognized
6	Error: <z> is not recognized
7	Error: <m> is not recognized
8	Error: <v> is not recognized
9	Warning: replacing invalid character code 131
10	Warning: discarding invalid character code 143
11	Warning: unescaped& which should be written as &amp
12	Warning: unescaped& or unknown entity "&:p"
13	Warning: unescaped& or unknown entity "&X"
14	Warning: <p> unexpected or duplicate quote mark
15	Warning: <p> missing '>' for end of tag
16	Warning: <j> missing '>' for end of tag
17	Warning: discarding unexpected <j>
18	Warning: unescaped& or unknown entity "&qT"
19	Warning: unescaped& or unknown entity "&Us"
20	Warning: discarding unexpected <bw>
21	Warning: unescaped& or unknown entity "&oX"
22	Warning: <I> attribute "4" lacks value
23	Warning: <I> missing '>' for end of tag
24	Warning: discarding unexpected <I>

4.5 EXPERIMENT AND RESULTS

In this section we present the results of the experiments we have carried out to compare our approach to other techniques in the literature. We describe the dataset used in our experimental study in section 4.4. We compare our approach to other approaches which are commonly used in extracting HTML pages in the literature. We performed our experiments on a

machine that was equipped with an Intel Core i3 processor that run at 2.40 GHz, had 2 GB RAM, Windows 7 pro 64 bit.

In our approach, we used the idea of Sleiman & Corchuelo [123] in 2013, they used TEX algorithm to extracts the information in the form of TextSet. The result of the extraction process is always a collection of TextSets which are labelled with computer generated labels. TEX does not translate the news articles into DOM trees. In our work, we translate the news articles into Tag tree. We analyse the performance of our approach using the three parameters precision, recall and F_1- measure.

4.5.1 OTHER APPROACHES

We compare our technique with other existing techniques.

ECON [122]: It takes a collection of news web pages and use HTML parser to create DOM tree. There is a node that wraps the entire contents of news with its subtrees, such node is known as summary-node. ECON finds a snippet-node which is the descendent of the summary-node. When snippet-node is found, then backtracks it until a summary-node is found, by which firstly wrapped the part of the news content, then backtracks from the snippet-node until a summary-node is found, and the entire content of news can be extracted after removing noise from the summary-node.

CoreEx [119]: They extract the main article from news web pages by using DOM tree where every node in the tree represents the HTML node of a web page. They score every node based on two counts, textCnt and linkCnt which means the amount of text and number of links it contains. Their algorithm runs on 1120 news web pages.

TEX [121]: It works on a collection of web documents. According to them input web documents not required to be translated into DOM Tree. Their approach works on malformed web documents without modifying them, and does not require the relevant information to be formatted using repetitive structures inside a web document. They works on two or more web documents and compares them in an attempt to discover shared patterns that are not likely to provide any relevant information, but parts of the template used to generate the web documents.

We take the idea of TEX in this work but we use Tag tree to discover shared patterns for the extraction of relevant information. The extract algorithm in TEX expand TextSet into three additional TextSets algorithm in Textset for extract algorithm while we only use shared occurrences by converting input news page into Tag tree by tokenization. In findpattern algorithm in TEX, they assume a shortest non–empty base Text of size s in TextSet ts and find a pattern inside base Text s that occurs in every Text in ts. In this work, to find matching pattern we use the observation that news web pages share the similar structure information having a particular alignment and forms similar pattern. We form the Tag tree of the structural information of the news web page, all leaves in a Tag tree share common Tokens. To find similar pattern we examine the path of the Tag tree to determine whether they are maximal similar or not. In filter algorithm in TEX, they filter out those patterns which show no variability in the TextSet. In our approach for filter out the undesired patterns two measure compactness and variability are used.

The target of ECON and CoreEx is similar to our approach; both the approaches extract the main contents from news web pages. However, the underlying algorithms of both the approaches are subsequently different. According to [119] [122], both ECON and CoreEx perform well, therefore we made a comparison among ECON, CoreEx , TEX, and the proposed approach.

4.5.2 PERFORMANCE ANALYSIS

Experiment results are often evaluated by Precision (P), recall (R) and F_1-measure (F1). We first run ECON and CoreEx on the same dataset in order to learn extraction rules, we then computed Precision, recall and F_1-measure. For each website we recognized one type of template. Then we manually analyse URL's of these pages to identify repetitive patterns such as URL structure, Home page, subsections etc. A regular expression is written for each website to match the template with high interest to us. This way we handcrafted annotations for every web document in our dataset that is used to calculate precision, recall and accuracy of our proposed approach. We could find which extracted ContentSet was the closest to each annotation. For the validation of our approach we compared each extracted ContentSet to every annotation.

The precision of a given category of dataset is the fraction of web pages of its computed category that are also found in the corresponding annotated category of the dataset. The recall of a given category is the fraction of web pages from the corresponding annotated category of dataset that were extracted from the same annotated category. To calculate these measures, we assign two or more web pages to the same categories if and only if they are similar. A true positive (TP) decision assigns two structurally similar web pages to the same category; a true negative (TN) decision assigns two structurally different web pages to different categories. A false positive (FP) decision assigns two structurally unlike web pages to the same category. A false negative (FN) decision assigns two structurally similar web pages to different category. Then the precision (4.2), recall (4.3), and F1-measure (4.4) are calculated as follows:

$$\Pr ecision(P) = \frac{TP}{TP + FP} \tag{4.2}$$

$$\mathrm{Re}\, call(R) = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 - measure(F1) = 2 \times \frac{P \times R}{P + R} \tag{4.4}$$

Table 4.3: Comparison of all Four Approaches

Category	**Proposed Approach**			**ECON**			**CoreEx**			**TEX**		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F1**
Business	0.97	0.96	0.96	0.82	0.54	0.65	0.80	0.85	0.82	0.96	0.95	0.95
Cricket	0.99	0.99	0.99	0.94	0.91	0.92	0.88	0.90	0.88	0.96	0.98	0.98
India	0.95	0.94	0.94	0.92	0.88	0.89	0.92	0.94	0.92	0.93	0.92	0.94
Tech	0.96	0.96	0.96	0.88	0.67	0.76	0.90	0.95	0.92	0.92	0.95	0.95
National	0.98	0.99	0.98	0.83	0.83	0.83	0.78	0.81	0.79	0.89	0.91	0.89
Science & Environment	0.93	0.86	0.87	0.76	0.77	0.76	0.85	0.67	0.74	0.92	0.83	0.85
Politics	0.98	0.99	0.93	0.67	0.67	0.67	0.60	0.55	0.57	0.97	0.98	0.97
World	0.95	0.96	0.95	0.91	0.88	0.89	0.83	0.83	0.83	0.93	0.73	0.82

Entertainment	0.96	0.94	0.94	0.73	0.68	0.70	0.71	0.78	0.74	0.92	0.93	0.93
Sports	0.92	0.87	0.89	0.64	0.76	0.69	0.55	0.64	0.59	0.84	0.78	0.81

Table 4.3 shows the P, R and F1 results of all three approaches ECON, CoreEx and our proposed approach when run on all 500 news pages. The result of the extraction process is always a collection of ContentSet. Experimental results show that the performance of our approach is higher than the other three approaches.

For the category Business, the values of P, R, and F1 of the proposed approach are better than the ECON, CoreEx, and TEX. In all the three approaches, ECON shows the lowest P (0.82), R (0.54), and F1 (0.65) values compare to the CoreEx (0.80, 0.85, 0.82), TEX (0.96, 0.95, 0.95), and the proposed approach (0.97, 0.96, 0.96).

For the category Cricket, CoreEx shows lowest values for the P (0.88), R (0.90), and F1 (0.88), compared to ECON 0.94, 0.91, 0.92, TEX 0.96, 0.98, 0.98, and the proposed approach 0.99, 0.99, 0.99.

For the category India, the P value of ECON and CoreEx show the same value 0.92, which has been lower than the TEX 0.93and the proposed approach 0.95. For the R and F1 values ECON show the lowest value 0.88 and 0.89 respectively, as compared to CoreEx (0.94, 0.92), TEX (0.92, 0.94) and the proposed approach (0.94 and 0.94).

For the category Tech, ECON shows lowest P, R, and F1 values as 0.88, 0.67, and 0.76 respectively, other than the three approaches. Proposed approach shows the highest value 0.96 for the all three P, R, and F1 values.

For the category National, CoreEx show the lowest P, R, and F1 values as 0.78, 0.81, and 0.79 respectively, than the other approaches. Our proposed approach shows the highest P, R, and F1 values as 0.98, 0.99, and 0.98 respectively.

For the category Science & Environment, the lowest value of P has been shown by the ECON 0.76 than the CoreEx (0.85), TEX (0.92), and proposed approach 0.93. The value of R has been shown lowest for the CoreEx 0.67 than the ECON (0.77) and TEX (0.83) while highest

for the proposed approach 0.86. For the F1 value, CoreEx shows the lowest value 0.74 than the ECON (0.76), TEX (0.85), and the proposed approach 0.87.

For the category Politics, CoreEx shows the lowest value than the other three approaches, While the proposed approach show the highest P, R, F1 values as 0.98, 0.88, and 0.93 respectively.

In the category World, proposed approach shows the highest P, R, and F1 values as 0.95, 0.96, and 0.95 respectively. While CoreEx shows the lowest value of P, R, and F1 as 0.83 for each.

For the category Entertainment, the lowest P value has been shown by the CoreEx 0.71, compare to the ECON, TEX, and proposed approach as 0.73, 0.93 and 0.96 respectively. For the value R, ECON, shows the lowest value as 0.68 compare to CoreEx, TEX, and proposed approach as 0.78, 0.93 and 0.94 respectively. For the F1 value, again ECON shows the lowest value 0.70 compare to the other three approaches CoreEx, TEX, and the proposed approach as 0.74, 0.93, and 0.94 respectively.

In the category Sports, for the values P, R, and F1, CoreEx show the lowest values 0.55, 0.64, and 0.59 comparison to ECON 0.64, 0.76, 0.69; TEX 0.84, 0.78, 0.81; and the proposed approach 0.92, 0.87, and 0.89 respectively.

From the above analysis, we can say that our proposed approach perform well than the other three approaches ECON, CoreEx and TEX. The performance of TEX is better than the ECON and CoreEx, and close to the proposed approach; as discussed in section 4.3, that we used the idea of TEX but constructs Tag Tree for pattern matching and filtering and improves the performance.

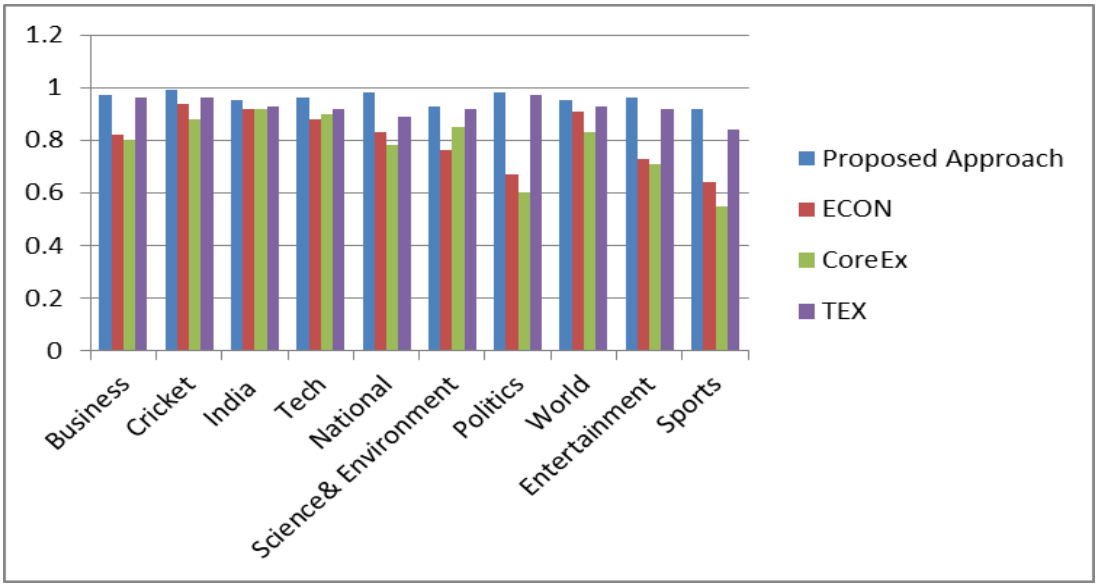


Figure 4.8. Precision Comparison of all Four Approaches

Figure 4.8, shows the graphical representation of comparison for all four approaches in terms of precison. Graph shows that the precision value of proposed approach is better than the other three approaches.

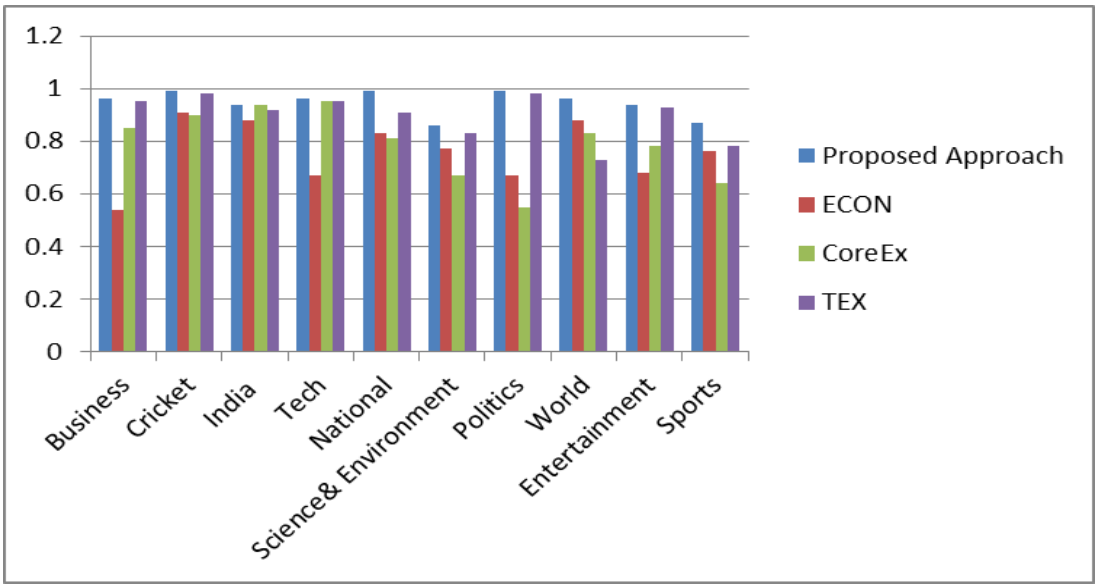


Figure 4.9. Recall Comparisons of all Four Approaches

Figure 4.9 describes the comparative analysis of the performance of all four approaches in terms of recall. It is clearly observed that for all ten categories of news articles, proposed approach shows highest recall values.

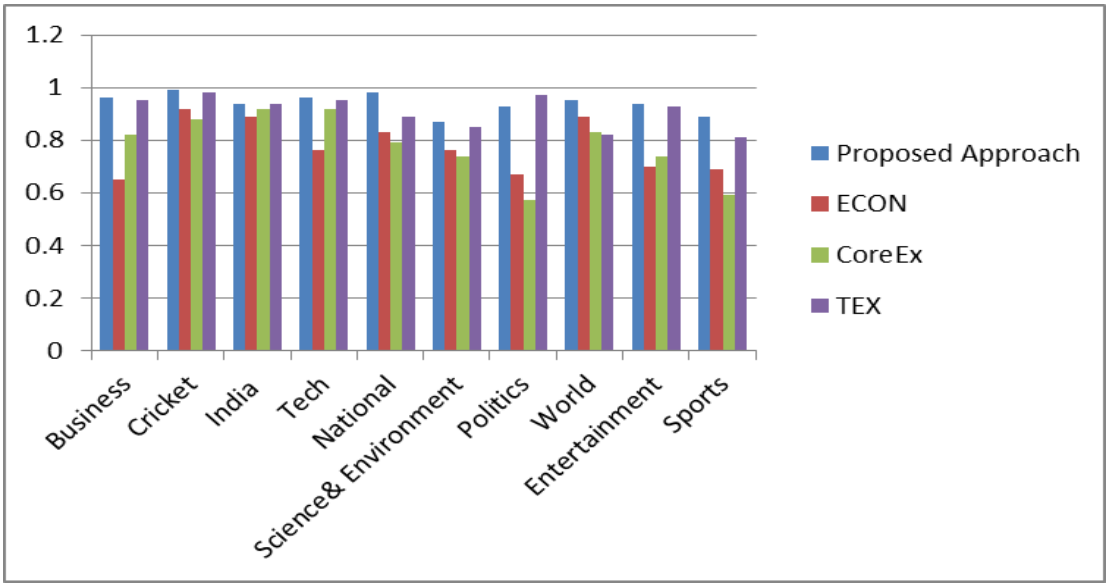


Figure 4.10. F1-Measure Comparison of all Four Approaches

A Figure 4.10 defines the comparative analysis of the performance of all four approaches in terms of F1- measure. Graph shows that the F1-measure of proposed approach is highest among the four approaches while ECON shows the lowest value.

4.6 SUMMARY

In present work, we presented a news web page content extraction approach to extract content from news web pages. Our approach applied to web news pages written in English. A news web page content extraction based on tag tree is proposed to efficiently extract meaningful information including records and data schema. In particular we have addressed the problem of finding and fetching news available on websites and extracting the relevant content. Through experimentation with ten news websites (Indian), we have demonstrated that our approach is highly effective for the task of content extraction.

Chapter 5

Keyphrase Extraction of News Web Pages

CHAPTER 5

KEYPHRASE EXTRACTION OF NEWS WEB PAGES

5.1 INTRODUCTION

The task of news web pages filtering and summarization requires the extraction of important keyphrses from the news document. Keyphrase extraction from news web pages is an important task for news documents summarization. This chapter discusses the proposed approach of keyphrase extraction from news web pages and performed by identifying candidate phrase from the news document and then calculates weight of the candidate phrase using various features and highest scorer candidate phrases are extracted as a keyphrases.

5.2 KEYPHRASE EXTRACTION

Under the growth of worldwide networking through the internet, the news consumption pattern moved from the traditional physical newspapers to online news aggregate system. As thousands of web news is posted on the internet every day, it is difficult to retrieve and summarize the relevant document effectively. So keyphrase extraction technique is used to provide the main phrases of a given web page. It is useful in many areas like summarization, automatic indexing, topic search and clustering [179]. Keyphrase extraction is one of the most important tasks in news web pages summarization. Readers make benefit from keyphrase because they can judge more quickly whether the news web page is worth reading. A summarization system tries to identify significant information that is important enough to be in the summary. In order to identify important topics and sentences in the documents, summarization system extracted keyphrases from the document.

Keyphrases are like index terms that enclose the important information about document content. Keyphrases actually offer concise and precise description of document content. Key phrases are considered as a single word or a combination of more than one word that represent the important concepts in a text documents. Document can be treated as a set of phrases; any

phrase in a new document can be extracted as a keyphrase. Phraseness and informativeness are the two main features of keyphrase. Phraseness is a fairly dynamic idea which depicts the degree to which a given word sequence is considered to be a phrase. Informativeness denotes how well a phrase catches or outlines the important notions in a set of documents. A set of keyphrases related to a document gives high-level description of a document content that helps readers in searching for relevant information.

Keyphrase extraction in a news web page has been a challenging research topic in recent years because news changes very rapidly. Only a small number of news websites have author given keyphrases and manually allocating keyphrases for each web news document is very effortful. Thus it is absolutely necessary to propose an approach for keyphrase extraction. Keyphrases of a document should be semantically related with the other words of the document. Therefore, in this work, we proposed a Keyphrase Extraction approach, which identifies the candidate keyphrases from documents and chooses those candidate keyphrase having highest weight score. Weight formula combines the feature set that includes TF*IDF, phrase distance in documents, and lexical chain is used for semantically related words that are interconnected by semantic relations. The number of words and the number of semantic relations among the words can be different for each lexical chain. WordNet [180] is used for the construction of lexical chain which is discussed in detail in section 5.4.4.

5.3 DESCRIPTION OF DATASET

The online news articles have been chosen from the ‘The Hindu’ news website. All these selected news is world news posted from 20 April 2016 to 30 April 2016. Our dataset contains 150 web news documents. The key purpose, we select ‘The Hindu’ news website for the experiment is that every news web page have author assigned keywords which are used to calculate the precision and recall values. We have taken the author assigned keywords as gold standard keyphrase. Some keyphrases have been chosen manually for each document. Most of the keyphrases consists of one or more than one words.

Keyphrases having more than three words are less in number in our dataset. Average number of manually assigned keyphrases per document is 15. Here it is interesting to note that all

author allotted keyphrase for a document may not occur in the title of the document. Total number of noun phrases in our dataset is 2250. The total number of author assigned keyphrases for all the documents in our dataset is 479.

5.4 PROPOSED APPROACH

In the proposed approach, firstly the document words are segmented, stemmed and stop words are removed. After that candidate phrases from the document are identified. Weight of each candidate phrase is computed by the features TF*IDF, phrase distance, and building lexical chain. According to the weight, a high scorer candidate phrases is selected as a keyphrases. The process of keyphrase extraction is shown in Figure 5.1.

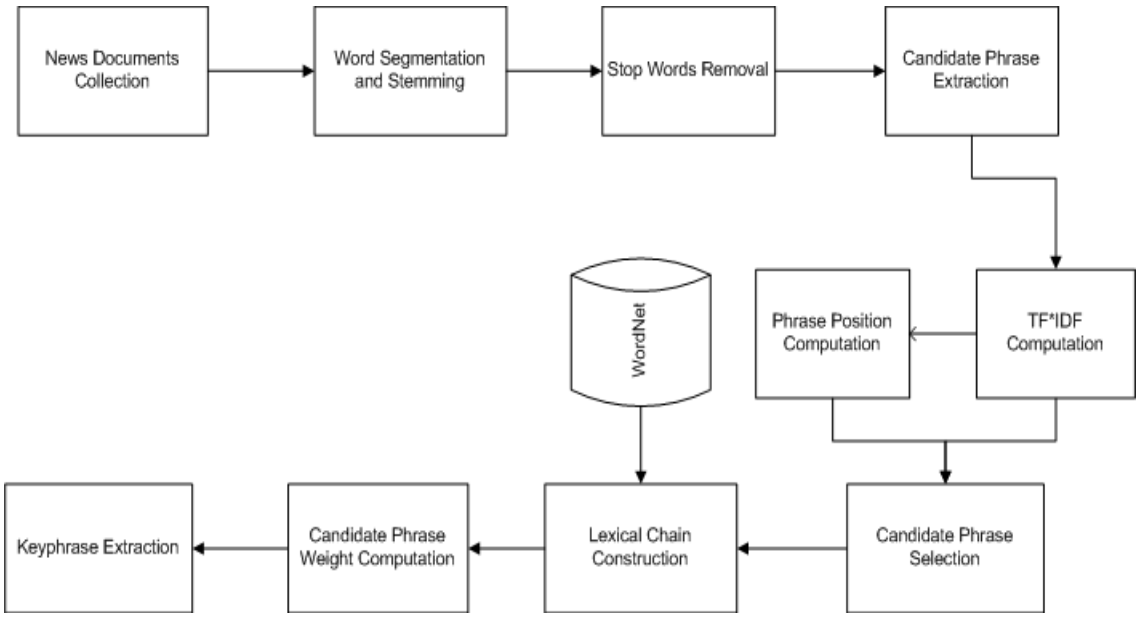


Figure 5.1. Flow Diagram of Keyphrase Extraction Process

The steps of the proposed approach are as follows:

Step 1: Words are segmented, stemmed and stop words are removed.

Step 2: Identify the candidate phrase from each document.

Step 3: Compute the TF*IDF and phrase distance of each candidate word

Step 4: Select the top n candidate phrases according to the value of TF*IDF and phrase distance.

Step 5: Build the lexical chains of each top n candidate phrase.

Step 6: Compute the weight of each candidate phrase.

Step 7: Select the top m candidate words as the keyphrase according to their weights. Select those candidate words as keyphrases which have higher weights.

In step 4 of the proposed approach, 'top n' represents the candidate phrases with high TF*IDF and phrase distance value. After weight computation of 'top n' candidate phrases, in step 7, 'm' represent the selected high scorer candidate phrases from 'top n' which we have denoted as keyphrases.

5.4.1. IDENTIFICATION OF CANDIDATE PHRASE

Keyphrases are extracted from candidate phrases. The noun phrases in the document are treated as the candidate keyphrase [6]. In order to recognize the noun phrases documents have been tagged by Stanford Part-Of-Speech (POS) tagger [181]. We used Stanford POS tagger to extract the lexical information about the terms in a document. Figure 5.2 shows the lexical tag assigned by the tagger for a document. According to this Figure, JJ, DT, NN, NNS, VBZ, NNP, PRP$, VBN, IN, CD, etc. are lexical tags assigned by the POS tagger.

Candidate keyphrase extracted from Figure 5.2 are: terrorists, central forensic science laboratory, DNA sample, government officials, National Investigation agency, investigation team, spokesperson, photographs and sensors.

Three|CD months|NNS after|IN Pathankot|NNP airbase|NN attacked|VBD terrorists|NNS belonging|VBG Pakistan-based|JJ Jaish-e-Mohammad|JJ (|NN JeM|NNP)|NNP forensic|JJ report|NN established|VBD six|CD terrorists|NNS present|JJ forensic|JJ examination|NN samples|NNS collected|VBN Billet|NNP last|JJ leg|NN operation|NN conducted|VBD found|VBN samples|NNS belonged|VBD "two|CD possible|JJ extract|NN any|DT "DNA|NN samples"|NN debate|NN number|NN terrorists|NNS present|NN airbase|NN came|VBD under|IN attack|NN alleged|VBD JeM|NNP terrorists|NNS intervening|VBG night|NN January|NNP 1|CD 2.|CD bodies|NNS terrorists|NNS killed|VBD within|IN first|JJ 24|CD hours|NNS operation|NN found|VBD airbase|NN bodies|NNS other|JJ found|NN Airmen's|NNS Billet|NNP where|WRB hiding|VBG blown|IN National|NNP Security|NNP Guard|NNP (|NNP NSG|NNP)|NNP different|JJ samples|NNS collected|VBN Billet|NNP Central|NNP Forensic|NNP Science|NNP Laboratory|NNP (|NNP CFSL|NNP)|NNP Chandigarh|NNP "The|NNP forensic|JJ report|NN says|VBZ samples|NNS tested|VBN positive|JJ human|NN remains|VBZ The|DT remains|NNS collected|VBD different|JJ rooms|NNS report|NN says|VBZ belong|JJ different|JJ humans|NNS remains|NNS badly|RB charred|VBN possible|JJ extract|NN DNA|NNP samples|NNS said|VBD senior|JJ government|NN official|NN National|NNP Investigation|NNP Agency|NNP (|NNP NIA|NNP)|NNP send|NN two|CD reminders|NNS CFSL|NNP before|IN reports|NNS sent|VBN February|NNP 8|CD The|DT Hindu|NNP giving|VBG details|NNS operations|NNS published|VBD interview|NN NSG|NNP Director-General|JJ R.C|JJ Tayal|NNP said|VBD certain|JJ six|CD terrorists|NNS airbase|NN sensors|NNS a|DT listening|NN device|NN wall|NN Airmen's|NNS Billet|NNP intercepted|VBD chatter|NN terrorists|NNS contacted|VBN Tuesday|NNP Mr.|NNP Tayal|NNP said|VBD "I|JJ seen|NN forensic|JJ report|NN always|RB said|VBD six|CD terrorists|NNS spokesperson|NN NIA|NNP said|VBD "|JJ want|NN comment|NN NIA|NNP preserved|VBD bodies|NNS four|CD terrorists|NNS shared|VBD photographs|NNS Pakistan|NNP through|IN letter|NN rogatory|NN Special|JJ Investigation|NN Team|NNP Pakistan|NNP expected|VBD visit|NN India|NNP conduct|NN joint|NN investigations|NNS

Figure 5.2. POS Tagged Document

Meanings of these tags are shown in Figure 5.3.

CD: Cardinal number, DT: Determiner, NN: Noun (Singular or mass), VBZ: Verb (3rd person singular present), To: to, VB: verb (base form), VBN: Verb (past participle), RP: Particle, IN: Preposition, NNP: Proper Noun, NNS: Noun (Plural), CC: Coordinating Conjunction, VBG: Verb (gerund), WP: Wh- pronoun, JJR: Adjective, Comparative, JJS: Adjective, Superlative, NNPS: Proper Noun (plural), PRPS: Possessive pronoun, VBP: Verb (non-3rd person singular present)

Figure 5.3. Meanings of the Tags

5.4.2. TF*IDF OF CANDIDATE PHRASE

After identifying candidate phrase, the collection of candidate phrases identified in the web news documents may be huge in number. From a vast collection, a small number of phrases may be selected as the keyphrases. In this work, we randomly selected some keyphrases from a single document. TF*IDF of each candidate phrase is used to rank the phrases. TF*IDF measure the phrase frequency in a document compared to its infrequency in general use.

We compute the TF*IDF of each candidate phrase by the given equation (5.1)

$$TF * IDF = \frac{t_f}{t_n} * \log(\frac{N}{n_i}) \qquad (5.1)$$

Where t_f is the frequency of term‘t’ in a document, ‘t_n’ is the total number of terms in a documents, ‘N’ is the total number of documents and ‘n_i’ is the number of documents in the dataset that contains term ‘t’.

5.4.3. PHRASE DISTANCE

The distance attribute is the position where a phrase first appears in the document. The candidate keyphrases that appear early in a document should be given higher score. Like previous approach [179] discussed in chapter 2, distance of a phrase from the beginning of a document is measured as the number of words that precede it initially seems divided by the number of words in the documents. The distance of a phrase in the document is calculated as in equation (5.2)

$$Phrase\,Dis\tan ce = \frac{n_j}{n} \tag{5.2}$$

Where ‘n_j’ is the number of words that come before its first appearance, and number of words in the document are denoted by ‘n’.

5.4.4. CONSTRUCTION OF LEXICAL CHAIN

Morris and Hirst [182] have first given the concept of Lexical Chain (LC). According to them lexical cohesion is an arrangement of related words that give the continuity of lexical meaning. Lexical cohesion occurs as a result of semantic relation between words. One of the main advantages of lexical cohesion is that it is an easily recognizable relation that enables the computation of lexical chain. Lexical chains visualize the semantically related words or phrases in the text. These words or phrases are called the lexical items and each item gives a specific meaning to a lexical chain. In this work, we use WordNet for creating lexical chains. With the help of path between concepts, lexical chain can be found. In general two concepts can have many possible lexical chains. For creating lexical chains we ignore numbers, units, currencies, times/periods, names, places and referring items [183]. For the construction of lexical chain we used synonym, hypernym/hyponym, coordinate term and meronym, Silber and McCoy [184] and Ercan [142] also used the same relations except coordinate term. In order to rank lexical chains, high scoring chains must be picked as the important concept from the original document. Further

we used Barzilay [185] idea of strong chain. It is defined as two words connected by a WordNet relations like Synonym and Hypernomy.

Figure 5.4 shows the different set of lexical chains chooses from the tagged document shown in Figure 5.2.

LC_1= {terrorists, terrorists attack, Pakistan-based JeM, JeM}
LC_2= {Central forensic science laboratory, forensic examination sample, forensic report, DNA sample}
LC_3= {Government officials, National Investigation Agency, Investigation Team}
LC_4= {Spokesperson, Photograph, Sensors}

Figure 5.4. Set of Lexical Chain

Lexical chain is created by taking a new phrase and finds a related chain for it according to lexical cohesion. For each new candidate phrase, the meaning of the phrase is examined. If the meaning is not matched with any of the existing chains, a new chain is created for this phrase and meaning is associated with this new chain. Otherwise, the phrase is connected to the matching chain.

From the Figure 5.4, we can quickly recognize that these phrases are related with each other. Lexical chains are formed of senses of phrase occurrences, not senses of unique words in the text. Each phrase in the lexical chains represents its intended sense of that candidate phrase.

Lexical chains usually depend on semantic relations that can be acquired from WordNet [180]. It is an online lexical reference system developed at Princeton University. Its taxonomy contains around 100,000 terms and grouped Noun, Verbs, Adjectives and Adverb into synonym set (synsets). The synsets are structured into senses based on the different meaning of the same term or concept. The synsets or concepts are related to other synsets by different type of relationships. The most common relationships are Hypernym/ Hyponym, Synonym/ Repetition, Meronym/ Holonym, Antonym, and Sibling relations used to build lexical chains. WordNet could be seen as the case of a semantic network that represents knowledge in the form of interconnection between synsets and the relations. Various representational techniques related to semantic networks assign numerical values or weights into relations [186] [187] [188].

According to Siblini and Kosseim [189] WordNet's 26 semantic relations do not contribute equally to the semantic relatedness between words. They manually analyse the relations and rank them according to their contribution like synonym gives highest contribution and get high rank while hypernym, sense, gloss, part, instance and other give lower contribution and get lower rank. The weights were simply assigned as the cost of traversing an edge or relation hence a lower weight is assigned to a highly contributory relation. Here, the used assigned weights are depends on the depth of its synsets in the WordNet taxonomy therefore rather than using the idea of Siblini and Kosseim [189], we give highly contributory relations higher weight. Figure 5.5 shows the lexical graph of LC_1 in detail.

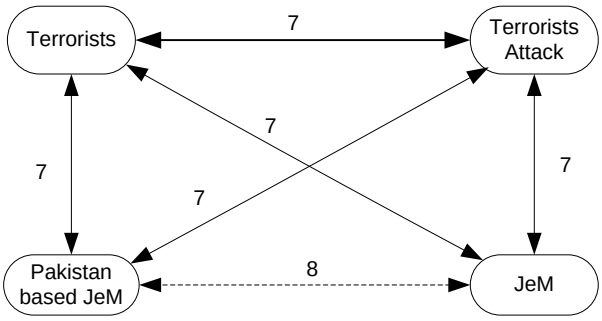


Figure 5.5. Lexical Graph

In the Figure 5.5, there are five Hypernym/Hyponym relations as {Terrorists $\overset{7}{\longleftrightarrow}$ Terrorists Attack; Terrorists $\overset{7}{\longleftrightarrow}$ Pakistan based JeM; Terrorists $\overset{7}{\longleftrightarrow}$ JeM; Terrorists Attack $\overset{7}{\longleftrightarrow}$ JeM, Pakistan based JeM $\overset{7}{\longleftrightarrow}$ Terrorists Attack} and one coordinate term defined as {Pakistan based JeM $\overset{8}{\longleftrightarrow}$ JeM} Weights of every relation between word senses are given intuitively [142]. Table 5.1, shows the allocated weights for the relation. Subsequent to scoring each lexical chain of the word, we select the chain with a maximum score as the lexical chain.

Table 5.1: Weight of Lexical Chain Relation

Relation	Explanation of Relation	Weight
Synonym/ reiteration	Same meaning	10
Coordinate Term	Sibling	8
Hypernym/hyponym	General/specific	7
Meronym	Is a part of	4

According to these assigned weights, the score of lexical chain LC_1 is equal to 43 (=5*7 + 8) since there are five Hypernym/Hyponym relations and one Coordinate term.

5.4.5. WEIGHT OF CANDIDATE PHRASE

Weight of a candidate phrase can be obtained by the combination of the features: TF*IDF, phrase distance, and lexical chain shown in equation (5.3).

$$Weight = a \times TF * IDF + b \times Phrase\,Dis\tan ce + c \times Lexical\,Chain \quad (5.3)$$

Where TF*IDF is the value of the candidate phrase, phrase distance is the distance of candidate phrase when it first appeared, and lexical chain is the length of the chain that contains candidate phrase and *a, b, c* are the parameters that can be adjusted. The value of these parameters in our experiment has been set to 1.

As an example, to compute the candidate phrase "*Terrorist*", selected from the dataset, the value of each component has been obtained as follows.

For computation of TF*IDF for the phrase "*Terrorist*", it is found in the document of dataset shown in Figure 5.2 at 7^{th} position and the number of words in the whole document is 367, then the value of TF*IDF is calculated as in equation (5.1)

$$TF * IDF = \frac{t_f}{t_n} * \log(\frac{N}{n_i})$$

TF*IDF= 0.0174832

Secondly we find the value of Phrase distance of the "*Terrorist*" phrase as calculated in equation (5.2)

$$Phrase\ Distance = \frac{n_j}{n}$$

Where the value of n_j is 6 in the document and n is 367. Therefore the value of phrase distance is 0.0163.

In the next step, construct and then calculate the value of lexical chain. We choose lexical chain LC_1 from Figure 5.4, because the value of LC_1 is higher than other lexical chains. The value of lexical chain LC_1 is 43, calculated in previous section 5.4.5. Finally the weight of the candidate phrase is calculated as:

W= 1*0.0174832+ 1* 0.245 + 1* 43

W= 43.0174832 ~ 43.01

Like the "terrorist" phrase, all the candidate phrases of the dataset are calculated. We then select top fifteen candidate phrases as keyphrases of a document.

5.5 EXPERIMENT RESULT AND EVALUATION

Experiments were carried out to evaluate the overall performance of our approach. For evaluating the automatically generated keyphrases, we first take the two standard information retrieval metrics precision and recall. The precision; measures the proportion of number of extracted key phrases that are also author tagged key phrases to the total number of extracted keyphrases. The second one 'recall' measures the proportion of the extracted keyphrases that are also author tagged keyphrases to the number of author tagged keyphrases. These metrics show how well generated phrases match a set of relevant phrases.

$$Precision = \frac{N_{e \cap t}}{N_e} \tag{5.4}$$

$$Recall = \frac{N_{e \cap t}}{N_t} \tag{5.5}$$

Where 'N_e' is the number of keyphrases extracted, 'N_t' the number of keyphrases tagged by author. $N_{e \cap t}$ is the number of extracted keyphrases that are also keyphrases tagged by author.

Table 5.2 shows the keyphrases assigned by the author of the news article which is the document number 2 in our dataset.

Table 5.2: Author Assigned Keyphrases for News Article Number 2 in the Dataset

Document No.	Author Key
2	Pathankot attack
2	forensic attack
2	Terrorism
2	Special Investigation Team
2	Joint investigation

From the document 2, our proposed approach extracted the top 5 keyphrases as shown in Table 5.3.

Table 5.3: Top 5 Keyphrases Extracted by Proposed Approach

Document Number	Author Key
2	Pathankot attack
2	forensic science laboratory
2	Terrorism
2	Special Investigation Team
2	National Investigation agency

Table 5.2 and Table 5.3 show that out of 5 keyphrases extracted by our approach, 3 keyphrases (Pathankot attack, Terrorism, Special Investigation Team) matched with the author assigned keyphrases. We also extract 10 and 15 keyphrases from the same document and results are shown in Table 5.4.

In order to compare our approach with state-of-the-art keyphrase extraction systems we have selected KEA [130] and KESR [170]. Most existing systems identify candidate phrases by the method applied in KEA and KESR.

KEA is comparatively simple and useful in automatic keyphrase extraction. The KEA identifies candidate keyphrase using lexical methods and calculates the feature value of each candidate phrase, and then predicts the good keyphrase from candidate by using machine learning algorithm. The basic model of KEA involves two stages. Firstly build a model for recognizing keyphrases by using training documents where the author keyphrases are known. Secondly, use the model create on first stage, choose the keyphrases from a new document. The

overall performance of KEA show that on average KEA can match between one and two of the five keyphrases chosen by the average author in the collection.

NFAS system considers all phrases except stop words in the web news pages. This system used Keyphrase Extraction based on Semantic Relations (KESR) approach for keyphrase extraction. The goal of KESR is to extract those words that have a low frequency but provide a major impact to the text subject. The basic model of KESR involves two attributes: TF*IDF, and word similarity and lexical chain. Word similarity is computed through HowNet. Extracted keyphrases compared with the phrases in the news title and phrases in the core hints provided by the author. By comparing their results with TF*IDF and KELC (Keyphrase extraction based on lexical chains) [190], KESR outperforms the other two in both the cases, when the title kept and when the title removed and core hints kept.

In this work, despite of calculating the value of each candidate phrase, select some top candidate phrase according to the value of TF*IDF and phrase distance and then construct the lexical chain of these selected candidate phrases using WordNet. Based on the all these features weight is computed and extracted the top scoring keyphrases.

We compare the overall performance of our proposed keyphrase extraction approach with the existing keyphrase extraction approaches. In the experiment, the number of keyphrases to be extracted was set to 5, 10, and 15 respectively. Table 5.4, shows that our approach shows better results than other approaches in terms of precision and recall.

Table 5.4: Precision and Recall Comparison of Three Approaches

Number of Keyphrases	**Average Precision**			**Average Recall**		
	Proposed Approach	KESR	KEA	Proposed Approach	KESR	KEA
5	0.34	0.32	0.28	0.25	0.24	0.29
10	0.22	0.20	0.19	0.46	0.36	0.40
15	0.17	0.18	0.15	0.51	0.41	0.48

Figure 5.6. Average Precision Comparison of Three Approaches

Figure 5.6 shows the comparison of the individual performance of three different approaches. Precision is the proportion of the keyphrases extracted that are correct. The experiments indicate that the precision of our approach when extracting 5 keyphrases is 0.34 which is 6.25% greater than KESR (0.32) and 21.4% greater than KEA (0.28). For extracting 10 keyphrase the precision of our approach is 0.22 which is 10% greater than KESR (0.20) and 15.7% greater than KEA (0.19) and finally for extracting 15 keyphrases the precision value of our approach (0.17) shows 5.6% lower value than KESR (0.18) while shows 13.3% higher results than KEA (0.15).

Figure 5.7. Average Recall Comparison of Three Approaches

Figure 5.7 shows the recall comparison of three different approaches. Recall is the fraction of relevant instances that are retrieved. Recall value of proposed approach, KESR and KEA when extracting 5 keyphrases is 0.25, 0.24 and 0.29 respectively where, proposed approach shows higher improvement in recall values as 4.2% than KESR and 13.7% than KEA. When extracting 10 keyphrases the recall value of our approach is 0.46 which is 27.8% higher than KESR (0.36) and 15% higher than that of KEA (0.40). Recall value of all the three approaches when extracting 15 keyphrases are 0.51, 0.41 and 0.48 respectively whereas for the proposed approach shows 24.4% greater value than KESR and 6.25% greater value than KEA.

5.6 SUMMARY

In this chapter, we have presented keyphrase extraction approach and candidate phrase detection as a part of keyphrase extraction where document is POS tagged first. It is implemented using TF*IDF, phrase distance and lexical chain. Experiment shows better results for the keyphrase extraction task. We have evaluated the approach with the parameters precision and recall.

Chapter 6

System Architecture of News Web Page Filtering and Summarization

CHAPTER 6

SYSTEM ARCHITECTURE OF NEWS WEB PAGE FILTERING AND SUMMARIZATION

6.1 INTRODUCTION

This chapter presents the complete architecture of the proposed news filtering and summarization system, the phases are as news web page classification, content extraction, and keyphrase extraction. In the previous chapters we have already discussed news web page classification, content extraction, and keyphrase extraction. Extracted keyphrases are important for sentence selection, selected sentences have been ranked using sentence weight and similarity measure is computed in order to reduce redundancy to obtained summarization.

6.2 NEWS SUMMARIZATION

News summarization is the task of creating a summary of one more news articles. Summarization systems take one or more documents as input and attempt to produce a concise and fluent summary of the most important information in the input.

Generate summary from multiple documents has gained interest since 1990s, most applications being in the domain of news articles. Several news summarization systems were inspired by research on multi-document summarization, for example Google News, Columbia News Blaster, or NewsInEssence.

A good summary is a readable one or it can be say that the sentences in the summary should be related to each other. On the other hand, they should be related and similar to the topic, and discuss the same information regarding the particular topic and this indicates that the summary should cover the original document content.

The goal of summarization is to present the most important information of the original document into a concise form while keeping its main content. It helps the user to quickly

understand the large volume of information and helps the reader quickly determine what the document is about and avoid reading the document itself.

In this work, we focus on extractive summarization which is a very robust method for document summarization it has been discussed in chapter 1. We assume that a user has access a stream of news stories that belong to similar subject; however the stream flows rapidly enough that nobody has the time to look at every story. In this situation, a person would choose to be kept up-to-date on the subject, and go through the details only when the reported topic is interesting enough therefore the system needs to gather news articles from various sources and link articles describing the same topic hence we used the multi-document summarization. One of the major problems in the multi- document summarization task is the identification of similarities and differences across documents we resolved this problem by using cosine similarity measure.

6.3 PROPOSED APPROACH AND SYSTEM ARCHITECTURE

This section describes the framework (as shown in Figure 6.1) of the proposed system in detailed which comprising four features discussed in next sections. The input is a collection of documents which are classified to news and non-news web pages in the news web page classification phase. The system extracts the news article content from the news web pages in the next phase. Each document covers one or more keyphrases and tries to pick sentences that cover keyphrases with respect to summary length. Then it extracts significant and non-redundant keyphrases in order to select sentences from news document. This phase generates set of sentences containing keyphrases. Now the weight of each sentence (discussed in section 6.3.4.5) is computed which is used for sentence ranking. In next step, redundancy is reduced by similarity computation using cosine similarity discussed in section 6.3.4.7. After eliminating the redundant sentences we select the final sentences for summary.

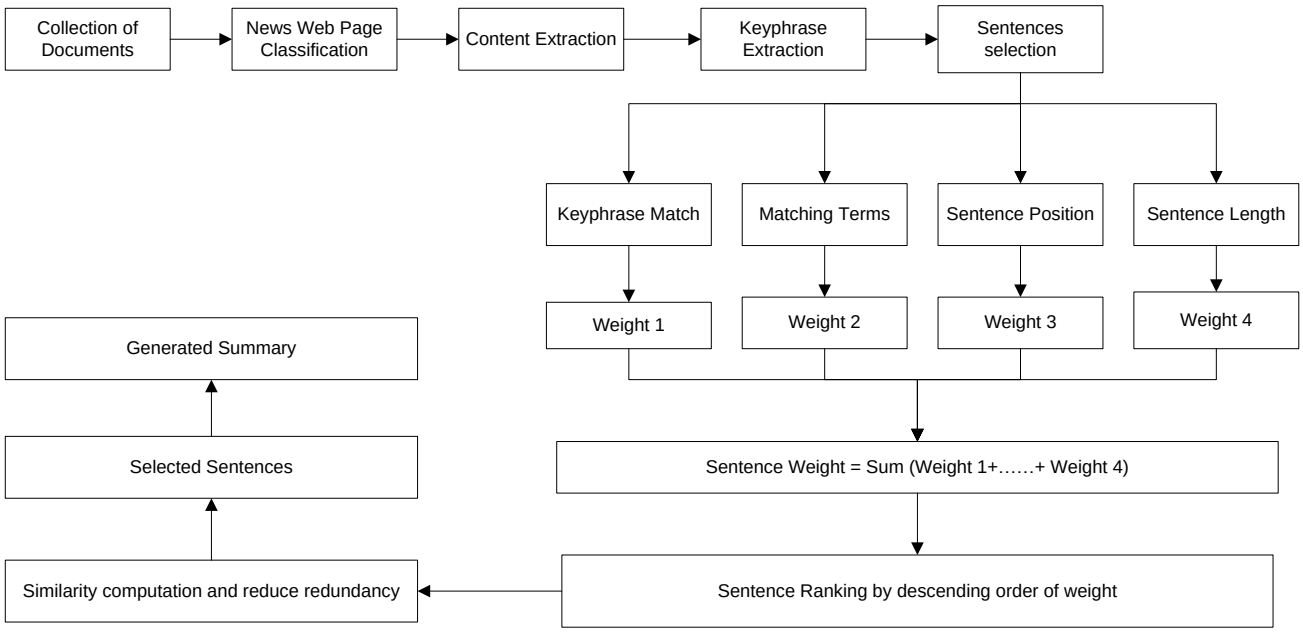


Figure 6.1. Proposed Architecture of News Filtering and Summarization

6.3.1 NEWS WEB PAGE CLASSIFICATION

In the work of news web page classification we propose an automatic recognition method that uses classification rules for web news based on a combination of content, structure and uniform resource locator (URL) attributes using Naïve Bayes algorithm discussed in detail in chapter 3. This phase classifies a news web page from a non-news web page. Correctly classified news web pages are used for the content extraction task.

6.3.2 EXTRACTING ARTICLE CONTENT

For extracting article content the approach is discussed in chapter 4. In the previous work for extracting content from news web pages we used the concept of tokenization of HTML pages; Web pages from different websites are parsed into Tag Tree and generated a template from each web page to discover matching patterns and multiple sequence alignment. In order to remove shared token sequences from the web pages until the relevant information is extracted from them. Extracted content are further used for keyphrase extraction and sentence selection.

6.3.3 KEYPHRASE EXTRACTION

Keyphrases are extracted from the sentences of the extracted content of news articles. The approach we used extract keyphrases from the candidate phrases where noun phrases of a document are treated as the candidate phrase. From these candidate phrases some top candidate phrases are selected by using TF*IDF and phrase distance. Then lexical chains are built from these selected candidate phrases and calculate weight. The high scoring candidate phrases are extracted as the keyphrases. The detailed approach is discussed in chapter 5. These extracted keyphrases are key elements in our summarization and used for the sentence selection.

6.3.4 SENTENCE SELECTION

Many web pages have diverse content, so it does not make sense to summarize the entire page as one unit. Rather, we believe it is best to select the sentences from the articles which are more significant. To generate a summary, highly ranked sentences are selected which are different from each other and cover the article main content with less redundancy.

Our goal was to find the sentence rank when making summaries of news articles. We use five kinds of features for sentence selection and ranking, including the direct keyphrase match, matching term, sentence position, sentence length, and to reduce redundancy we use cosine similarity model.

6.3.4.1 Direct Keyphrases Match

Keyphrases are used to evaluate the sentence importance. After extracting the set of keyphrases for each document, our goal is to pick sentences for each document that cover more important and non-redundant keyphrases. Essentially, keyphrases that have been repeated in more sentences are more important and could represent more important keyphrase. Therefore sentences that contain more frequent keyphrases are more important. The approach of keyphrase extraction we discuss in previous section 6.3.3 and chapter 5.

For determined which of the number of keyphrases (N) are more important than others in the summary. We calculated the keyphrase frequency by normalizing the count of keyphrases ki by the count of all keyphrases Ki-N in the document set and non keyphrases ~ ki.
Qazvinian et al. [191] define the record of building a summary comprising a set of keyphrases S as in equation (6.1)

$$F(S) = |S \cap A| \tag{6.1}$$

Where 'A' denotes the set of all keyphrases from sentences which are not included in the summary. We tried to apply Qazvinian's method and achieved poor results. This was because our data set is completely different from what Qazvinian was working with. Therefore we tried several different functions to select sentences.

We score the sentence by direct keyphrase match. Those keyphrases which occurs in two or more sentences are more important than others. We calculate the direct keyphrase match by the following formula shown in equation (6.2)

Direct Keyphrase match = When two or more sentences were contain same keyphrases.

$$K_{match} = \frac{K_N}{T(S)} \tag{6.2}$$

Where 'K_{match}' denotes the direct keyphrase match in the document set. 'K_N' denotes the number of times keyphrase occurs in the document set. 'T(S)' denotes the total number of sentences in the document set.

6.3.4.2 Matching Terms

When terms previous and next to the keyphrases are matched, an extra score is added to the sentence.

Table 6.1: Keyphrase Score

Keyphrases	Score of direct Keyphrase match	Number of matching terms
K. Srikant	38%	8
Australian open	14%	7
Chen Long	6%	3

Shuttler	8%	6
Olympic	6%	8
Badminton	3%	2

Table 6.1 shows the score of the keyphrase, we assume the threshold of 5 matching terms for the sentence selection, therefore the sentences above the threshold of 5 matching terms; we added an extra 1% score to the score of direct keyphrse match.

6.3.4.3 Sentence Position

We take the feature Sentence Position (SP) from the work of Wong et al. [92], is a simple and effective feature for the news summarization. The perception is that leading sentences in the news article contain summarizing information. We used the positional information, occurrence of a sentence in the document weather the sentence occurs very early or very late in a document boost the top sentences of an article shown in equation (6.3).

$$SP(S) = 1 - \frac{P}{N} \tag{6.3}$$

Where 'N' is the total number of sentences in the articles and 'P' is the position of the sentence 'S' in the article.

6.3.4.4 Sentence Length

Sentence Length (SL) [92] is a binary feature which helps in reducing the noisy short text in the summary. It checks if the sentence contains at least 10 numbers of words. The sentences below the given limit of 10 words will be ignored to generate summary, as shown in equation (6.4).

$$SL(s) = \begin{cases} 1 & if\ len(s) \geq 10 \\ 0 & Otherwise \end{cases} \tag{6.4}$$

6.3.4.5 Sentence Weight

According to the above four features, we compute the final significant score of a sentence by specifying a certain weight for each kind of feature, as shown in equation (6.5)

$$SentenceWeight = \lambda_1 Keyphrase\,match + \lambda_2\, Matching\,Terms + \lambda_3\, Sentence\,Position + \lambda_4 Sentence\,Length \quad (6.5)$$

Where λ_1, λ_2, λ_3, λ_4 represent the weight parameters for the four kind of features. The value of these parameters is lying between 0 and 1 according to the features. If keyphrase match is maximum or exact match then we set value '1' for λ_1. If matching terms are more than three then λ_2 is set to '1' otherwise '0.5'. If sentence position is higher or lies in top five than λ_3 is set to '1' otherwise '0.5'. If sentence length is greater than 10 then λ_4 is set to '1' otherwise set to '0'.

6.3.4.6 Sentence Ranking

After weight computation we use algorithm 1 for sentence ranking. It accepts Sentence Set (SS) as input and produce ranked list of sentences in descending order.

Algorithm 1

Input: Sentences in Sentence Set (SS)

Output: A list of sentences sorted by descending order of their weight (LS).

```
1. begin
2.          for each sentences S in SS
3.              if (length (S)<8)
4.                      Remove S into SS
5.               else
6.                       Add S in SS
7.                  ranking (SS, SenWeight)
8.            end for
9.       Output LS
10. end
```

In this algorithm, sentences are ranked in main loop run from (step 2 to 9). We have fixed the length of the sentence by 8 words and only those sentences which satisfy this condition are included in the Sentence Set (SS) to calculate the score otherwise it has been removed from the sentence Set.

6.3.4.7 Similarity Model For Redundancy Reduction

News summarization task face the major problem of identifies the similarity and differences across news articles. The reduction of similarity or redundancy is a complex task because the properties of individual sentences are dependent on other sentences included in the summary. The basic idea in our research is to find out the similarity of the news articles in the same event where several sentences may have a substantial information overlap. Some researchers used clustering to obtain group of similar sentences [192] while in this work to find similarity among sentences depends on keyphrases in the news articles that report about the same event and link the similar keyphrases together.

Summarizers can identify similarities and differences among documents by comparing and merging representations of document content from the analysis phase. Summarizers have been developed that can eliminate redundant information across stories to provide a concise summary.

The similarity measure is usually based on matching keyphrases only. We used keyphrase based similarity because different articles use different styles for writing the same event and articles sentences are not same line by line. In this step similarity among keyphrases in the news articles are found that report about the same events and link the similar keyphrases together. Our approach uses TF-IDF, phrase distance and lexical chains to identify the several keyphrases which convey approximately the same information discussed in the previous section 4. Find the similarity among the same event. We analyses the news stories written in English language. We have limited our focus to the textual contents of the articles. Thus pictures and other multimedia are rejected. News articles from different news websites described the same event in different aspects, user often compare articles from different sources. Therefore news articles are gathered from different news websites and link articles describing the same event. News article are updated frequently and their description are overlapped each other in series of news articles. Therefore by removing duplicate description user can obtain efficient information. We have to summarize different news articles on the same event in a single extract. It’s far from clear that sentence scores from different news articles should be comparable.

If the terms of two sentences are very similar, the sentences may probably have approximate feature values, therefore, they may also probably have approximate scores. Thus the extracted summary may include high score sentences which are very similar, this will cause redundant information in summary. Therefore, we need to remove sentences that are redundant to others in articles on an event. Minimize redundancy between passage and the selection of most representative sentences are important issues in summarization. For redundancy identification of news articles we use cosine similarity to measure the similarity between documents. Once similar passage in the input documents have been identified, the information they contain must be included in the summary. Our task is to assign a score to every sentence that indicates the importance of that sentence in the summary. We use algorithm 1 to select the sentences. The formula of cosine similarity is shown in equation (6.6).

$$\cos\theta = \frac{d1.d2}{|d1||d2|} \tag{6.6}$$

Our system collects news articles that need to be summarized. First of all article collection are spit into sentences in such a way that they are indexed by a letter and a number combination. The letter shows the corresponding document and number indicate the sentence position within its respective document [193].

Example 1: we collect the five news article from different sources that belong to the same event. We take one sentence from each document. We consider one sentence until the first full stop is occurred.

D1 s1= India shuttler Kidambi Srikanth defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.

D2 s1= The **Badminton** Association of India announced a cash award of Rs. 5 lakh to **Srikanth** for clinching the **Australian Open Super Series** title in Sydney on Sunday.

D3 s1= **Kidambi Srikanth** has resurfaced on Indian **badminton**'s horizon since the surprise of a win over Lin Dan in the China Open final in November 2014.

D4 s1= Indian **shuttler Kidambi Srikanth** notched up his second successive Super Series title with a stunning straight-game triumph over reigning Olympic champion **Chen Long** in the **Australian Open summit** clash in Sydney on Sunday.

D5 s1= **Kidambi Srikanth** is enjoying the best phase of his career.

Table 6.2 shows the five news documents first sentence matching according to keyphrases.

Table 6.2: Documents Matching According to Keyphrases

Documents	**Text**	**Keyphrases**
D1	India shuttler Kidambi Srikanth defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.	India, shutttler, Kidambi Srikant, Olympic champion, Chen Long, Austrlia open super series, title, sunday
D2	The **Badminton** Association of India announced a cash award of Rs. 5 lakh to **Kidambi Srikanth** for clinching the **Australian Open Super Series** title in Sydney on Sunday.	Badminton, India, Kidambi Srikant, Australian open super series, title, Sydney, Sunday
D3	**Kidambi Srikanth** has resurfaced on Indian **badminton**'s horizon since the surprise of a win over Lin Dan in the China Open final in November 2014.	Kidambi Srikanth, Indian, Badminton
D4	Indian **shuttler Kidambi Srikanth** notched up his second successive Super Series title with a stunning straight-game triumph over reigning Olympic champion **Chen Long** in the **Australian Open summit** clash in Sydney on Sunday.	Indain, shuttler, Kidambi Srikanth, Olympic, Chen Long, Australian Open, Sydney, Sunday
D5	**Kidambi Srikanth** is enjoying the best phase of his career.	Kidambi Srikanth, career

Similarities of documents are shown in Table 6.3.

Table 6.3: Similarity Measure of Documents

	D1S1	**D2 S1**	**D3 S1**	**D4 S1**	**D5 S1**
D1S1	1	0.71	0.43	0.93	0.38
D2 S1	0.71	1	0.65	0.8	0.38
D3 S1	0.43	0.65	1	0.41	0.59
D4 S1	0.93	0.8	0.41	1	0.36
D5 S1	0.38	0.38	0.59	0.36	1

First all the sentences are ordered by score from highest to lowest, and then the summary sentences are selected iteratively, each time the current candidate sentence is compared to the sentences already in the summary. If the sentence is not too similar to any sentence already in the Summary or if the similarity value of the two sentences is lower than a given threshold, then the sentence is selected for the summary.

Table 6.3 shows the similarity among the first sentence of five documents. In cosine similarity '1' denotes the exactly similar sentence; these sentences can cause redundancy in the summary, therefore removed from the final summary. Exactly different sentences are denoted by the '0' in the cosine similarity such type of sentences are also does not contribute in the summary and excluded from the final summary. Based on the example document similarity measure computed in Table 6.3, we set the threshold value of cosine similarity is 0.65. When two sentences are approximately similar (that is close to 1) the one with the higher weight is selected for the summary.

6.4 GENERATE SUMMARY

Summary is formed by extracting top ranking sentences according to scores assigned to the sentences. However, to reduce redundancy, we use cosine similarity model [192]. A sentence is selected for summary generation if it gets the highest rank and not too similar to any sentences existing in the summary. To determine similarity between sentences we use cosine similarity at threshold t = 0.65.

The following Algorithm 2 describes the summary generation strategy in our system; it uses the Algorithm 1 for sentence ranking.

Algorithm 2

Step 1: Extract the main news content from the documents.

Step 2: Extract the keyphrases from the news documents.

Step 3: Compute sentence weight as per the equation (5)

Step 4: Sort sentences in descending order of weight using algorithm 1.

Step 5: Remove redundant sentences based on cosine similarity.

Step 6: Final selected sentences are used in summary generation.

6.5 EXPERIMENTAL DATASET

In order to evaluate our work we collect the news articles from five different news websites named as The Economic Times, The Hindu, The Times of India, Hindustan Times, and Indian Express. Ten events which occurred between 15 may 2017 to 16 June 2017, were manually selected by these five news websites. Each event contained five articles, which were reported in the same day. The news events are selected from different categories as market, business, India, Tech, Nation, Science & Environment, politics, world, Entertainment, Sports. An annotator reads all the news articles and connects the keyphrases that discuss the same story. We discuss the dataset in detail in chapter 7.

6.6 SUMMARY

In this chapter, we describe the overall working of our proposed news summarization approach. Briefly, the news web pages are first classified non-news web pages using Naïve Bayes algorithm. After that important content are extracted from the classified news web pages. The news summarization then looks for the keyphrase extraction which extracts the important keyphrases from the extracted content. Further news summarization system select sentences according to the sentence ranking and reduce similarity by using cosine similarity measure.

Chapter 7

Results and Analysis

CHAPTER 7

RESULT AND DISCUSSION

7.1 INTRODUCTION

This chapter presents the result and discussion of our proposed news filtering and summarization system. The evaluation of news summarization method has been done on the dataset consisting of 100 news articles and for evaluation of the system ROUGE (Recall-Oriented Understudy for Gisting Evaluation) tool has been used to evaluate the quality of generated summaries by counting overlapping units between the candidate summary and the reference summary. The tool presents five ROUGE measures including ROUGE-N is a n-gram recall between a system summary and reference summary where 'n' denote the length of the n-gram, ROUGE-L measure the longest matching sequence of words using Longest Common Subsequence, ROUGE-W is the Weighted Longest Common Subsequence, ROUGE-S is skip-bigram co-occurrence statistics that measure the overlap of skip-bigrams between a system summary and a reference summary; and ROUGE –SU, it is the extension of ROUGE-S with the addition of unigram as counting unit. The higher the ROUGE score has, the better the system is. In this work, we used ROUGE-1, ROUGE-2 and ROUGE-SU4 for the evaluation task, because according to Lin [195] these measures worked reasonably well for multi-documents summarization. The chapter also illustrates a comparison between our approach and other baseline approaches.

7.2 EXPERIMENTAL DATASET

The experimental dataset consists of ten different categories of news articles namely Market, Business, India, Technology, National, Science & Environment, Politics, World, Entertainment, Sports. Each category has two sets and each set contains five news articles on the same topic related to the individual category i.e. total of 10 articles per news category so a total of 100 news articles have been included in the dataset. All news articles are collected from five

famous Indian news websites (English) such as The Economic Times, The Hindu, The Times of India, Hindustan Times, and Indian Express. Number of sentences contained in these news articles range from 10 to 60. We manually constructed the 150 word summary as reference summary. Details of this dataset are given in Table 7.1.

Table 7.1: Analysis of Dataset

S. N.	Topic	Number of article set	Number of articles in both sets	Average number of Sentences per article set	Average number of words per article set
1	Market	2	10	42	464
2	Business	2	10	36	514
3	Sports	2	10	49	593
4	India	2	10	34	554
5	National	2	10	39	667
6	Technology	2	10	45	385
7	World	2	10	53	753
8	Politics	2	10	47	581
9	Entertainment	2	10	25	497
10	Science& Environment	2	10	37	209
11	Total	20	100	40	632

7.3 EXPERIMENTAL SETUP

ROUGE [193] has been used as an automatic evaluation method and it based on the similarity of n-gram. ROUGE has been tested for extraction based summaries with a focus on content overlap [196] [197] [198]. It is one of the standard ways to compute effectiveness of auto generated summaries by comparing it to a set of reference summaries that is typically produced by the human. There are several metrics within the ROUGE and most widely used are ROUGE-1, ROUGE-2 and ROUGE-SU4. ROUGE-1 and ROUGE-2 computes the unigram and bigram overlap between the computers generated and reference summaries, whereas ROUGE-SU4 calculates the skip bigram overlap with up to four intervening terms.

Short summaries obtained from the proposed approach have been evaluated and is also compared with ROUGE scores of reference summary as given in Table 7.2. Performance of the proposed approach for all the ten categories is shown in Table 7.3. Finally the comparative analysis with three baseline approaches on same dataset is also listed in Table 7.4, and performance improvements are shown in Table 7.5.

7.4 EVALUATION

Automated machine summaries can be compared with reference summaries (human summaries) using ROUGE summarization evaluation tool. It works by comparing an automatically generated summary against a set of reference summaries.

The input to the system is a collection of two sets of documents related to the same topic. The output is a concise set of two summaries providing the condensed information of the input documents. A good evaluation measure should assign a good score to a good summary and poor score to a bad summary.

Computation of ROUGE-1, ROUGE-2 and ROUGE-SU4 values for the system and reference summary sentences has been described in Table 7.2. Example sentences have been taken from the documents (D1, D2, D3, D4, and D5) of our dataset discussed in chapter 6.

Table 7.2: Example of System and Reference Summary Sentences

System Summary Sentence	Srikanth defeated Olympic champion Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.
Reference Summary Sentence	India shuttler Kidambi Srikanth defeated Olympic champion from China Chen Long in a dominating fashion to win the Australia Open Super Series men's singles title on Sunday.

According to the above example shown in Table 7.2, the values of ROUGE -1, ROUGE-2 and ROUGE-SU4 are computed as:

ROUGE-1 refers to the overlap of unigrams between the system summary and reference summary. In the above example there are 22 words in the system summary which matched with words of reference summary. The formula for determining the ROUGE-1 value can be demonstrated as follows:

$$ROUGE-1 = \frac{Matching\ unigrams\ in\ system\ and\ reference\ summary}{Total\ number\ of\ unigrams\ in\ reference\ summary} \quad (7.1)$$

ROUGE-1= 22/27= 0.815= 81.5%.

ROUGE -2 refers to the overlap of bigrams between the system summary and reference summary.

$$ROUGE-2 = \frac{Matching\ bigrams\ in\ system\ and\ reference\ summary}{Total\ number\ of\ bigrams\ in\ reference\ summary} \quad (7.2)$$

ROUGE-2 = 11/13 = 0.846 = 84.6%

ROUGE-SU4 is an extended version of ROUGE-2 that allows maximum 4 length word-level gaps between the bigram [4].

$$ROUGE-SU4 = \frac{SKIP\ 4\ (system\ summary, reference\ summary)}{reference\ summary, 4} \quad (7.3)$$

ROUGE-SU4= 5/6 = 0.833 = 83.3%

Further, to have testing of our proposed approach we take the articles from different categories to obtain the ROUGE values for each of these categories separately. Results are shown in Table 7.3.

Table 7.3: All Three ROUGE Values for Different Type of Categories

Categories		**Sentences**	**Words**	**ROUGE-1**	**ROUGE-2**	**ROUGE-SU4**
Market	**Set 1**	28	382	73.09%	72.03%	70.23%
	Set 2	42	537	84.21%	82.29%	85.53%
Business	**Set 1**	37	502	77.77%	79.94%	76.24%
	Set 2	39	522	83.13%	81.32%	79.67%
Sports	**Set 1**	34	457	73.27%	76.55%	72.54%
	Set 2	37	489	81.05%	80.78%	81.58%
India	**Set 1**	49	695	95.99%	92.67%	94.04%
	Set 2	43	549	90.34%	89.21%	91.55%
Technology	**Set 1**	18	277	69.11%	64.28%	65.79%
	Set 2	32	435	78.98%	74.69%	77.98%
National	**Set 1**	41	529	89.98%	88.57%	89.48%

	Set 2	36	472	77.26%	78.14%	76.57%
Politics	**Set 1**	43	553	91.89%	87.85%	91.09%
	Set 2	42	542	89.20%	85.34%	88.66%
World	**Set 1**	27	384	72.76%	71.53%	71.66%
	Set 2	35	463	86.72%	83.31%	85.98%
Entertainment	**Set 1**	46	569	94.32%	91.92%	93.38%
	Set 2	38	511	88.86	86.02%	87.66%
Science& Environment	**Set 1**	32	399	76.01%	79.23%	77.36%
	Set 2	21	192	70.41%	69.85%	66.76%

In the above table, first column contains the ten different categories of news articles (as already discussed). Second column contains the two set of news articles for each particular category and their corresponding sentences and words are given in column third and fourth respectively. Last three columns shows the ROUGE-1, ROUGE-2, and ROUGE-SU4 values respectively.

From the above results, we can say that better results were observed on the categories having large number of sentences and words size, and poor results are found for those categories where sentences and words are small in size. Like in India category, set 1 contains the highest value of sentences and words 49 and 695 respectively therefore shows the highest ROUGE-1, ROUGE-2 and ROUGE-SU4 values as 95.55%, 92.67%, and 94.04% respectively. While in category Technology, set 1 contains the lowest value of sentences 18 and words 277, hence shows the lowest ROUGE-1, ROUGE-2 and ROUGE-SU4 values as 69.11%, 64.28%, and 65.79% respectively.

7.5 COMPARATIVE EVALUATION OF PROPOSED APPROACH WITH OTHER BASELINE APPROACHES

There are various approaches are available, out of which we choose the three baseline approaches SRRank, TSES, and LAKE for comparison with our proposed approach. To know the accuracy of our proposed approach we used the same dataset for all the approaches. The brief description of baseline approaches is:

SRRank [159]: An extractive multi-document summarization system. It used semantic role information to enhance multi-document summarization and Saliency score of all sentences are obtained by greedy algorithms for sentence selection.

TSES [154]: This approach extracts important keyphrases to select the important sentences. Each sentence ranked according to the specified features and extracts the highest ranking sentence to generate the final summary. TSES generates summaries in four steps, firstly removes stop words and assigning POS tag for each word in the document. In the second step extract important keyphrases from the document and rank them by implementing a new algorithm. In the next step sentences are ranked according to the extracted keyphrases and in the final step reduced the amount of the candidate sentences in the summary in order to produce a qualitative summary using KFIDF measurement.

LAKE [150]: LAKE is a multi-document summarization approach for DUC-2005. This approach used the idea of keyphrase extraction as a useful approximation to summarization. It uses a machine learning framework to select significant keyphrases for a document. Summaries are generated considering both the relevance and the coverage of keyphrases for a certain topic.
The reason of using these approaches as our baseline is that, both LAKE and TSES used the keyphrase-based approach for their experiments and we have also keyphrase as an important feature in our approach. The SRRank incorporates the semantic role information into the graph-based ranking algorithm and we also used semantic role information for lexical chain construction.

In the proposed approach keyphrase is identified as an important feature for sentence ranking. For sentence ranking we used some different features than the baseline approaches like direct keyphase match, matching terms, sentence position and sentence length. Experimental results show that these combinations of features gives better results than other baseline approaches. Proposed approach is based on sentence ranking and minimization of redundancy in the summarization.

We already conduct the experiments for redundancy reduction which show that our approach is helpful for sentence ranking and to minimize redundancy for the news summarization. Experiments conducted in this chapter (Table 7.3) show that our news

summarization approach gives accurate and precise summary for the multi-document news articles. Further a comparative analysis of our results with other baseline approaches would help us understand the overall performance of the proposed approach with the other popular approaches i.e. SRRank, TSES, and LAKE. For the all three baseline approaches used the same dataset having ten different categories containing two set each.

Table 7.4 shows the ROUGE-1, ROUGE-2 and ROUGE-SU4 score for each set from ten different news websites.

From the table, results show that, in average, proposed approach seems to perform better than the other three baseline approaches for all ROUGE values.

Further, in Table 7.5 below the actual improvement in performance of our approach compare with each of the three baseline approaches for ROUGE-1, ROUGE-2, and ROUGE-SU4 values.

Table 7.4: Experimental Evaluation of Dataset

		ROUGE -1				ROUGE-2				ROUGE-SU4			
		Proposed Approach	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**	**Proposed Approach**	**SRRank**	**TSES**	**LAKE**
Market	**Set 1**	**73.09%**	69.28%	71.29%	74.34 %	**72.03%**	68.77%	69.11%	72.98%	**70.23%**	68.27%	69.93%	71.04%
	Set2	**84.21%**	79.95%	80.67%	82.47%	**82.29%**	79.86%	80.05%	81.03%	**85.53%**	79.98%	80.87%	83.18
Business	**Set 1**	**77.77%**	74.86%	76.24%	75.65%	**79.94%**	74.83%	75.89%	78.44%	**76.24%**	72.59%	73.99%	75.37%
	Set 2	**83.13%**	78.02%	79.78%	81.99%	**81.32%**	77.04%	77.94%	79.82%	**79.67%**	74.78%	75.67%	78.02%
Sports	**Set 1**	**73.27%**	71.64%	69.97%	72.99%	**76.55%**	73.86%	74.65%	75.23%	**72.54%**	70.75%	71.78%	71.98%
	Set 2	**81.05%**	77.87%	78.06%	80.56%	**80.78%**	76.01%	77.21%	78.35%	**81.58%**	76.24%	79.67%	80.99%
India	**Set 1**	**95.55%**	89.47%	88.88%	92.87%	**92.67%**	88.59%	89.12%	90.43%	**94.04%**	89.48%	88.97%	91.79%
	Set 2	**90.34%**	87.09%	89.17%	89.88%	**89.21%**	83.25%	85.87%	86.98%	**91.55%**	88.78%	86.08%	89.23%
Technology	**Set 1**	**69.95%**	65.52%	67.48%	70.15%	**64.28%**	62.82%	63.89%	65.21%	**65.79%**	63.15%	62.24%	66.03%
	Set 2	**78.98%**	74.03%	75.44%	77.89%	**74.69%**	72.02%	73.15%	74.06%	**77.98%**	73.36%	73.01%	76.97%
National	**Set 1**	**89.98%**	84.09%	85.32%	87.23%	**88.57%**	82.39%	83.66%	86.76%	**89.48%**	84.54%	85.39%	88.56%
	Set 2	**77.26%**	74.78%	75.89%	76.99%	**78.14%**	73.02%	74.88%	76.98%	**76.57%**	71.93%	72.78%	74.87%
Politics	**Set 1**	**91.89%**	85.45**%**	88.67**%**	90.54%	**87.85%**	82.51**%**	83.41**%**	85.98%	**91.09%**	87.99**%**	87.15**%**	89.79%
	Set 2	**89.20%**	84.77%	87.01%	87.96%	**85.34%**	81.82%	81.09%	83.75%	**88.66%**	83.86%	84.45%	87.77%
World	**Set 1**	**72.76%**	71.87**%**	71.98**%**	73.98%	**71.53%**	70.98**%**	69.02**%**	72.89%	**71.66%**	68.64**%**	69.54**%**	72.11%
	Set 2	**86.72%**	80.75%	81.83%	84.55%	**83.31%**	79.24%	78.58%	80.69%	**85.98%**	80.79%	82.65%	84.57%
Entertainment	**Set 1**	**95.99%**	89.09**%**	92.80**%**	94.34%	**92.67%**	87.45**%**	88.56**%**	90.05%	**94.04%**	89.11**%**	90.04**%**	92.29%
	Set 2	**88.86**	83.53%	83.79%	85.39%	**86.02%**	80.05%	81.88%	84.99%	**87.66%**	81.93%	83.21%	86.16%
Science& Environment	**Set 1**	**76.01%**	73.71**%**	74.04**%**	74.97%	**79.23%**	75.77**%**	74.78**%**	78.65%	**77.36%**	73.24**%**	73.96**%**	75.02%
	Set 2	**70.41%**	68.54%	69.94%	71.25%	**69.85%**	67.62%	68.12%	69.95%	**66.76%**	63.79%	64.52%	67.32%

Table 7.5: Performance Improvement of Proposed Approach over the Baseline Approaches

Categories		ROUGE -1			ROUGE-2			ROUGE-SU4		
		Improvement over SRRank	Improvement over TSES	Improvement over LAKE	Improvement over SRRank	Improvement over TSES	Improvement over LAKE	Improvement over SRRank	Improvement over TSES	Improvement over LAKE
Market	**Set 1**	5.5%	2.5%	-1.7 %	4.7%	4.2%	-1.3%	2.8%	0.42%	-1.1%
	Set2	5.3%	4.3%	2.1%	3.0%	2.8%	1.6%	6.9%	5.7%	2.8%
Business	**Set 1**	3.9%	2.0%	2.8%	6.8%	5.3%	1.9%	5.0%	3.0%	1.1%
	Set 2	6.5%	4.2%	1.3%	5.6%	4.3%	1.8%	6.4%	5.2%	2.1%
Sports	**Set 1**	2.3%	4.7%	0.38%	3.6%	2.5%	1.7 %	2.5%	1.0%	0.77%
	Set 2	4.1%	3.8%	0.61%	6.2%	4.6%	3.1 %	7.0%	2.3%	0.72%
India	**Set 1**	5.4%	6.1%	1.6%	3.7%	3.1%	1.6%	4.3%	4.9%	1.7%
	Set 2	3.7%	1.3%	0.51%	7.2%	3.9%	2.6%	3.1%	6.3%	2.6%
Technology	**Set 1**	6.7%	3.6%	-0.28 %	2.3%	0.61%	-1.4%	4.1%	5.7%	-0.4%
	Set 2	6.6%	4.7%	1.3%	3.7%	2.1%	0.85%	6.2%	6.8%	1.3%
National	**Set 1**	7.0%	5.4%	3.2%	7.5%	5.8%	2.1%	5.8%	4.7%	1.0%
	Set 2	3.3%	1.8%	0.35%	7.0%	4.3%	1.5%	6.4%	5.2%	2.2%
Politics	**Set 1**	7.5%	3.6%	1.5%	6.4%	5.3%	2.2%	3.5%	4.5%	1.4%
	Set 2	5.2%	2.5%	1.4%	4.3%	5.2%	1.8%	5.7%	4.9%	1.0%
World	**Set 1**	1.2%	1.1%	-1.6%	0.77%	3.6%	-1.9 %	4.3%	3.0%	-0.60%
	Set 2	7.3%	5.9%	2.6 %	5.1%	6.0%	3.2%	6.4%	4.0%	1.6%
Entertainment	**Set 1**	7.7%	3.4%	1.7%	5.9%	4.6%	2.9%	5.5%	4.4%	1.8%
	Set 2	6.3%	6.1%	4.0 %	7.4%	5.1%	1.2%	6.9%	5.3%	1.%
Science& Environment	**Set 1**	3.1%	2.6%	1.3%	4.5%	5.9%	0.73%	5.6%	4.5%	3.1%
	Set 2	2.7%	0.67%	-1.1%	3.2%	2.5%	-0.1%	4.6%	3.4%	-0.8%

From the above Table 7.5, for the category market, proposed approach shows better results over SRRank and TSES in set 1 for ROUGE-1, ROUGE-2 and ROUGE-SU4 values while LAKE shows better results than the proposed approach of 1.7%, 1.3% and 1.1% respectively for the all three ROUGE values. In set 2, proposed approach shows better results than all the three baseline approaches and highest improvement over SRRank in ROUGE-SU4 of 6.9%.

For the category business, the results of proposed approach are better than all the three baseline approaches SRRank, TSES and LAKE for the ROUGE-1, ROUGE-2 and ROUGE-SU4 values in both sets and the highest improvement is shown over SRRank by 6.5% in set 2 for the ROUGE-1 value while the lowest improvement is shown over LAKE in set 2 for ROUGE-SU4 value.

For the category sports, proposed approach shows better results than all the three baseline approaches in both sets for the ROUGE-1, ROUGE-2 and ROUGE-SU4 values, while the highest improvement is observed over SRRank in set 2 for the ROUGE-SU4 value and the lowest improvement is over LAKE 0.77% in set 1 for ROUGE-SU4.

For the category India, our approach shows better results compared to all the three approaches for ROUGE-1, ROUGE-2, and ROUGE SU4 values for both sets, while the highest improvement is shown over TSES by 6.3% for ROUGE-SU4 value in set 2 and the lowest improvement is shown by LAKE 1.7% in set 1 for ROUGE-SU4.

For the category Tech, in set 1 the performance of proposed approach is better than only SRRank and TSES while shows poor results than LAKE for all three ROUGE values. In set 2 proposed approach outperforms all the three baseline approaches and the highest performance is shown over TSES 6.8% in ROUGE-SU4.

For the category Nation, the results of proposed approach are better than all the three approaches SRRank, TSES and LAKE for ROUGE-1, ROUGE-2 and ROUGE SU4 values in both sets and we can see best result over SRRank by 7.5% for ROUGE-2 value in set 1.

For the category Politics, all the three approaches SRRank, TSES and LAKE shows poor results than the proposed approach in both sets for ROUGE-1, ROUGE-2 and ROUGE- SU4 values. The proposed approach shows highest improvement over SRRank by 7.5% for ROUGE-1 value in set 1 and the lowest improvement in set 2 by LAKE for ROUGE-SU4 value.

For the category world, for ROUGE-1 value our approach shows better results only from SRRank and TSES in set 1 while LAKE shows better results than the proposed approach for all the three values of ROUGE (ROUGE-1, ROUGE-2 and ROUGE- SU4) as 1.6%, 1.9% and 0.6% respectively. In set 2, proposed approach shows better improvement than all the three baseline approaches and the highest improvement is shown over SRRank 7.3% for ROUGE -1.

For the category Entertainment, proposed approach outperforms all the three approaches for ROUGE-1, ROUGE-2 and ROUGE- SU4 values in both set. The highest performance improvement is shown over SRRank in set 1for ROUGE-1 and the lowest improvement is shown over LAKE in set 2 for ROUGE-2.

For the category Science & Environment, in set 1, proposed approach shows better results than all the three approaches for ROUGE-1, ROUGE-2 and ROUGE- SU4 values, and we can see the highest improvement over TSES for ROUGE-2. In set 2, proposed approach shows better results only by SRRank and TSES while in comparison of LAKE, shows poor results for all three ROUGE (ROUGE-1, ROUGE-2 and ROUGE- SU4) values as 1.1%, 0.1% and 0.8% respectively.

Overall the proposed approach performs better than other baseline approaches. Among the three approaches LAKE is the strongest, and it can outperform the other two TSES and SRRank approaches. By the analysis of the results we can also say that the performance of proposed approach is affected by the size of the number of sentences and number of words. That category which contains large size of sentences and words shows good results otherwise shows poor results like in Market, Technology and World.

The graphical representation of all the ROUGE values for both set 1 and set 2 from Table 7.4, are shown in Figure 7.1 to Figure 7.6.

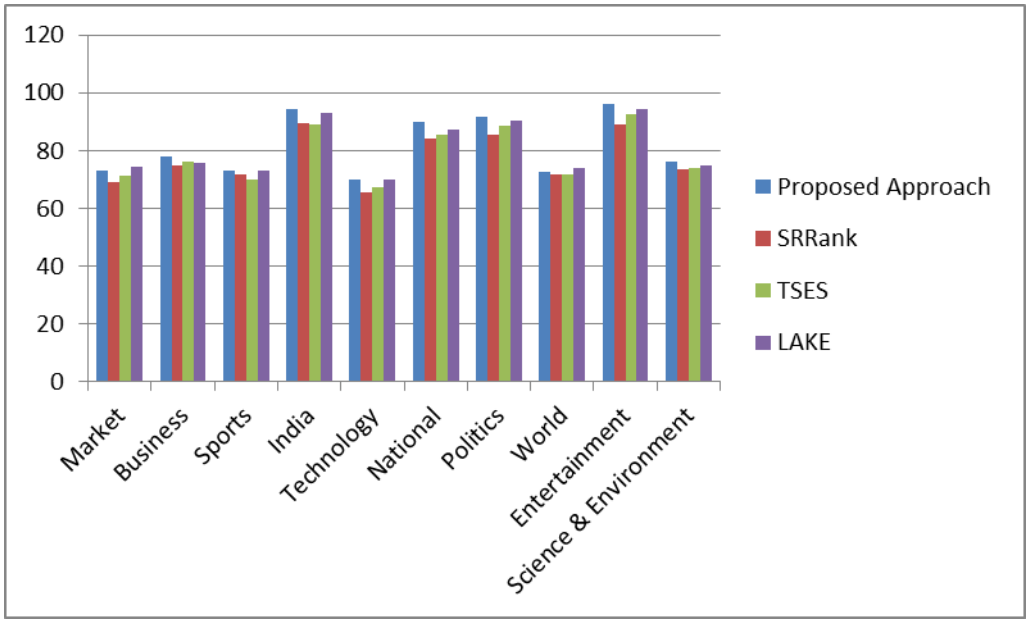


Figure 7.1. Comparison of ROUGE-1 Values for Set 1

Figure 7.1, depicts the graphical representation of the performance of all the approaches for the ROUGE-1 values in set 1. From the graph, we can say that proposed approach outperform over all the other three approaches, except in three categories market, tech, and world; LAKE shows better results than our proposed approach. Overall the proposed approach performs better in seven out of ten categories.

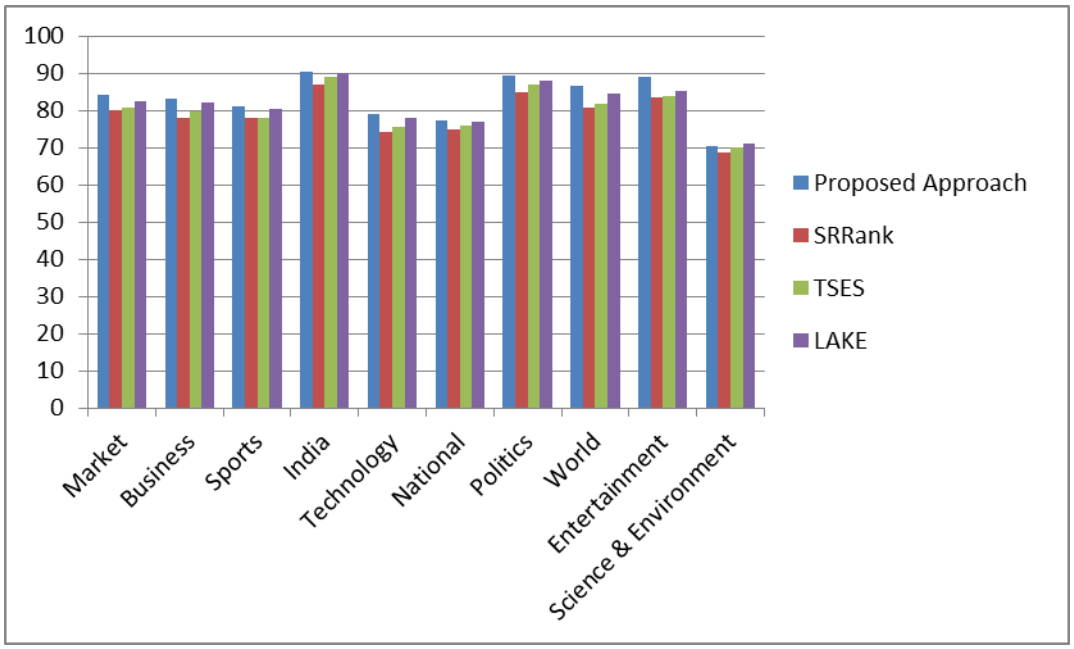


Figure 7.2. Comparison of ROUGE-1 Value for Set 2

Figure 7.2 shows the graphical representation of performance comparison of all the approaches for ROUGE-1 values in set 2. Graph shows that the performance of proposed approach is better than that of three baseline approaches in nine out of ten categories while in the category of Science & Environment shows poor performance.

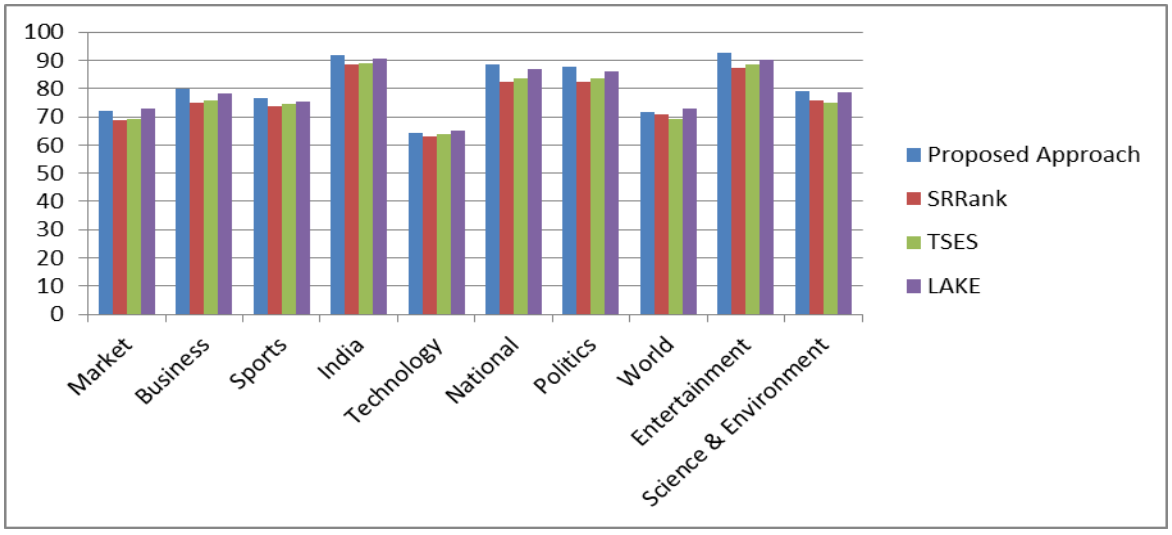


Figure 7.3. Comparison of ROUGE-2 Values for Set 1

For the ROUGE-2 values of set 1 Figure 7.3 shows the graphical representation of all approaches performance. Graph depicts that proposed approach shows good results in seven out of ten categories. Proposed approach performs poor in three categories compared to LAKE as Market, Technology, and world.

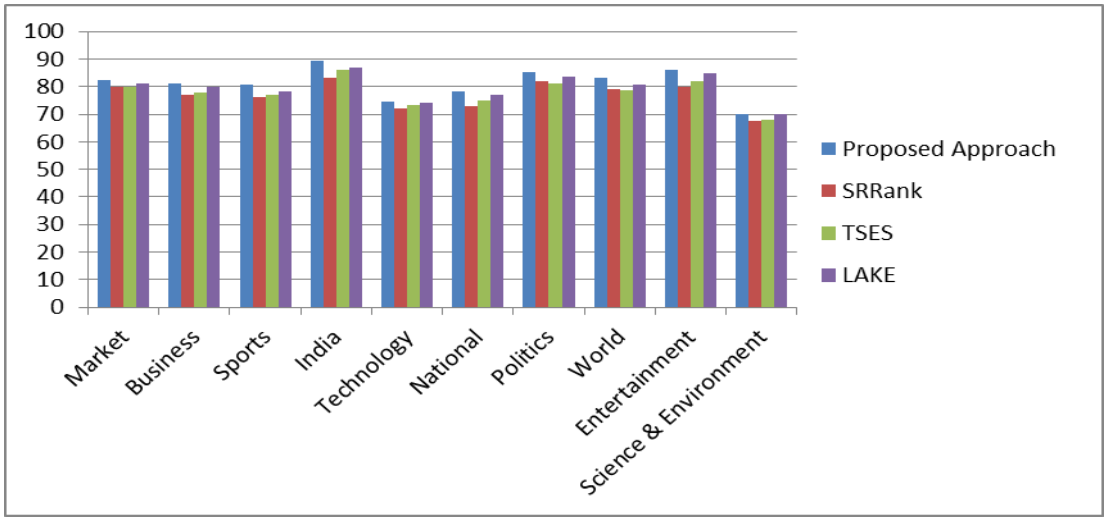


Figure 7.4. Comparison of ROUGE-2 Values for Set 2

Graphical representation of the performance of all the approaches of ROUGE-2 values for the set 2 is shown in Figure 7.4. Graph show that overall performance of proposed approach is better than that of baseline approaches. Proposed approach shows better results in nine out of ten categories while in science & Environment LAKE shows better results than our proposed approach.

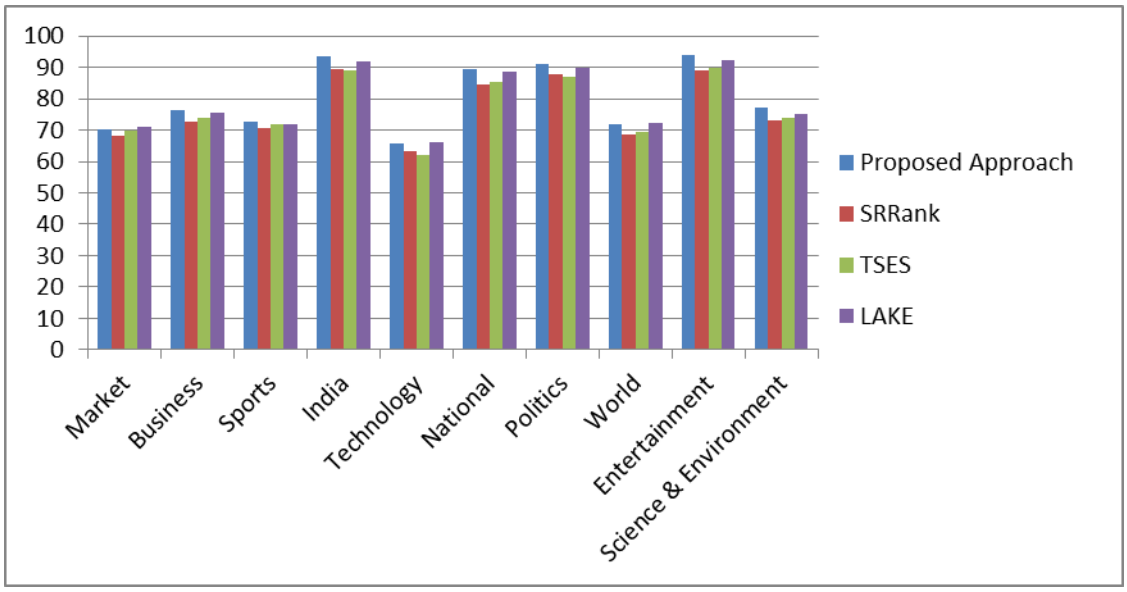


Figure 7.5. Comparison of ROUGE-SU4 Value for Set 1

Performance comparison results of all the four approaches for ROUGE-SU4 values of set 1 are shown in Figure 7.5. According to the above figure, we can say that overall performance of proposed approach is better than that of other baseline approaches. Our approach shows better results in seven out of ten categories while in three categories (Market, Technology and world) LAKE perform better than the proposed approach.

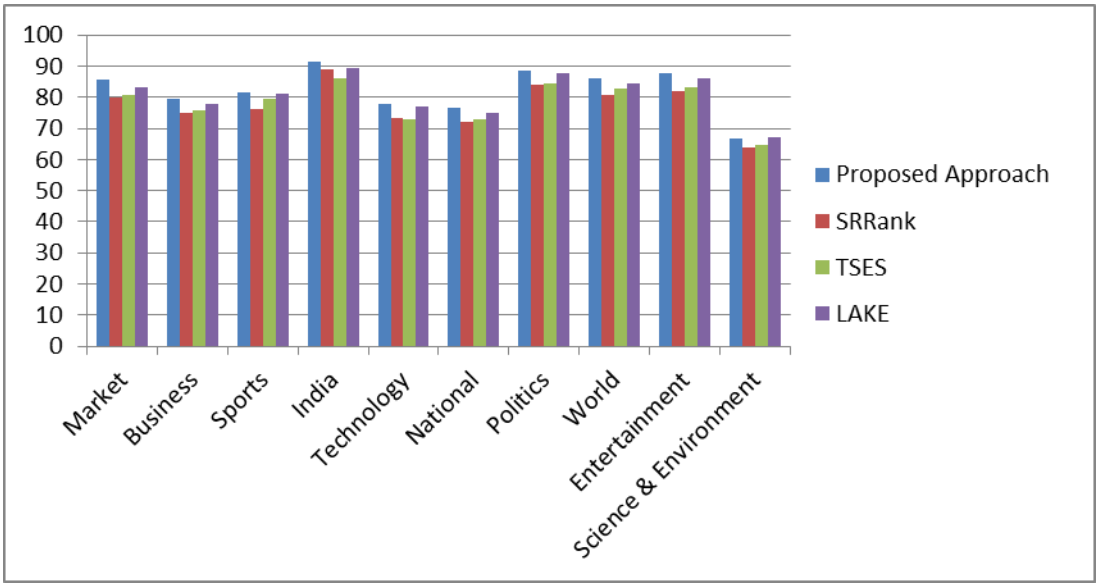


Figure 7.6. Comparison of ROUGE-SU4 Values for Set 2

The graphical depiction of the performance of all the approaches for ROUGE-SU4 values of set 2 are shown in Figure 7.6. According to the graph, proposed approach performs better than the other three baseline approaches. The performance of the proposed approach is better in nine out of ten categories while in only one Category (science & Environment) LAKE performing better than our approach.

7.6 SUMMARY

The overall analysis of this research states that keyphrase extraction and similarity measure is an effective technique that helps in precise summarization of news web pages. Ranking also plays an important role in sentence selection. From the experiment result, there can be seen a good ROUGE score of our proposed approach compared to the standard baseline news summarization approaches.

Chapter 8

Conclusion and Future Work

CHAPTER 8

CONCLUSION AND FUTURE WORK

With the huge amount of data available on web, the research in area of news web page summarization has turn out to be very useful for users. Recently, there have been significant advantages in the area of news summarization. With the emergence of erroneous amount of online data, it is desirable to construct a news summarization approach that can extract, compare and rank sentences to create a summary of various news articles.

We have made our effort for developing a news filtering and summarization system for Indian news websites (English). News web page filtering and summarization has been increasingly gaining more attention from the research community. There have been few existing systems developed for news summarization but little effort has been done on the combination of supervised algorithms based classification, content extraction and keyphrase extraction for summarization. To achieve this task we have gone through various phases which involve news web page classification, content extraction, keyphrase extraction and finally summarization using various features for sentence selection and ranking.

Initially, our effort has been to take a comprehensive overview of types of news summarization and extractive and abstractive summarization approaches. We also discussed all categories of text classification and observed that the supervised text classification is better than other approaches.

A news web page classification phase classifies a news web page from a non-news web page. Prior knowledge of correct news page allows us to narrow our content extraction task considerably. Therefore, news web page classification is considered as most important task in the news filtering and summarization.

The growing domain of online newspaper presents a rich area which can benefit greatly from automatic classification approach. An effective approach has been presented to realize the automatic news web page classification based on content attributes, structure attributes and URL

attributes of news web pages. We begin with the observation that the proper choice of attributes can have a significant impact on the performance of classification algorithm. Attributes has been extracted from the ten different news websites, and used Naïve Bayes algorithm for classification and conducted comparative experiments with various existing algorithms on the same dataset for the both set of experiments, and the results show that Naïve Bayes perform better than other algorithms as 11% better precision value than SMO and 24% better precision value than J48, for the first set experiment. While in second set of experiment, Naïve Bayes shows 14% better precision value than SMO and 18% better precision value than J48.

Next phase of our proposed news filtering and summarization system is the content extraction. The process of content extraction is defined as the extraction of relevant content from massive data, such as text, database, semi-structured and multimedia documents. Efficiently extracting high quality content from news web pages is a challenging yet important problem in the field of information retrieval and news summarization. We presented a news web page content extraction approach to extract content from news web pages. It is based on the idea that web documents from different websites share similar tokens, and these tokens generate Tag tree to efficiently extract meaningful information that saves end users from the burden to learn extraction rules and maintain extraction rules. The content of the news web page extracted by finding similar patterns and filter out those patterns which are not contain any useful information. Experiments showed that our approach improves existing results in the literature (Guo et al., 2010; Prasad and Paepcke, 2008) for the problem of content extraction from news web pages and perform extraction with high precision and recall. Our approach applied to web news pages written in English. A news web page content extraction based on tag tree is proposed to efficiently extract meaningful information including records and data schema. In particular we have addressed the problem of finding and fetching news available on websites and extracting the relevant content. Through experimentation with ten news websites, we have demonstrated that our approach is highly effective for these tasks.

In the next phase, the task of news web pages filtering and summarization requires the extraction of important keyphrses from the news document. Readers make benefit from keyphrase because they can judge more quickly whether the news web page is worth reading. We have presented an effective approach which can extract keyphrases from news web page.

Our approach takes noun phrases of the documents as a candidate Phrase, and used POS tagger for this task. While ranking candidate keyphrases, weights of each candidate keyphrase is measured and choose the highest score keyphrases. To determine the weight of the keyphrase we use the TFIDF, phrase distance in the document and lexical chain. The approach is evaluated by the evaluation parameters precision and recall. Experimental results show that this approach is competitive with other known approaches.

In the final phase of our work we presented an approach for extractive summary of multiple news articles based on keyphrsae-based sentence weight and use cosine similarity to reduce redundancy. To calculate the weight of the sentence we combine four features keyphrase match, matching term, sentence position and sentence length and to reduce redundancy we used cosine similarity. We compare our proposed approach with other approaches on English news documents dataset. All experiments results indicate that our approach performs well on several multi-document summarization approaches for English news documents.

8.1 RESEARCH CONTRIBUTION

Exploration of news filtering and summarization system is an important research area. Following are the major contribution of our research:

1. We have tried to develop a news filtering and summarization model for Indian news websites (English) with the intension to summarize news articles from different sources.
2. We create dataset from ten different Indian news websites for news web page classification and summarization.
3. News web page classification approach correctly classifies the news web pages from non-news web pages, Correct classification of news web pages are important for news summarization.
4. Content extraction based on Tag Tree efficiently extracts meaningful information from news articles.
5. Keyphrase gives brief and precise information about the article. Accurately extracted keyphrases plays an important part in news summarization. Therefore, our approach used lexical chain based keyphrase extraction and shows better results.

6. The selection of features for sentence selection and ranking is important for precise summarization. The suitability and appropriateness of the approach is reflected by the results.

8.2 FUTURE WORK

Our objectives for the future are ambitious; it remains to be seen how many of them are achievable. There are several possible extensions to this study in every phase.

Most news web pages have small comments section at the end of the page where readers post their opinions about news articles. The textual information of comments section not quite relates to the news article, therefore, problematic for the analysis.

In news summarization, since news articles may have been written at different times, for readers, it is important to make sure that what happened and when, in order to fully understand the news story and event. Reordering of sentences is not sufficient to accomplish this task. Therefore, in future it would be interesting to add temporal phrases such as “on Tuesday” or “three days later” that place the event described in a given sentence and gives the overall context of the summary that helps readers to understand the timeline of the event.

During this research work, we came across so many other aspects which though are not directly the task and scope of this research and may be taken in future to add dimension in this research. Some of the future work may include:

1. Our future target is to explore a method to detect and eliminate the noisy information of discussion boards.
2. There is a possibility to improve our approach to extract the content also from blog and forum pages. Since the blog pages and forum pages have similar characteristics as news web pages.
3. In keyphrase extraction, when author assigned keyphrases are less than the automatic extracted keyphrases then it is difficult to assess the automatic extraction results. Hence there will be accomplice more research on searching for a more logical and target approach to assess the automatic extraction results.

4. This research can be extended to include Multi-Lingual news articles by using the proposed approach in this work.
5. The dataset used in the present work can be extended to larger dataset for the more confident results.
6. Proposed approach can also be applied to some other datasets to test its robustness.
7. Also, we will try to utilizing more evaluation methods to evaluate the proposed summarization approach.

References

REFERENCES

1. DARPA – Defense Advanced Research Projects Agency, Unites States Department of Defense, http://www.darpa.mil.
2. BT'sProsumhttp://www.ontoweb.org/workshop/amsterdamdec8/ShortPresentation/njkamsterdamdec 2000.pdf.
3. Luhn, H.P, "*The Automatic Creation of Literature Abstracts",* In IBM Journal of Research Development 2(2), pp. 159–165, 1958.
4. Edmundson, H.P, "*New Methods in Automatic Extracting",* In Journal of the Association for Computing Machinery 16(2), pp. 264–285, 1969.
5. Jones, K. S., "*What might be in a summary*?", Information retrieval, 93, 9-26, 1993.
6. Bawakid, A., & Oussalah, M, "*A Semantic Summarization System: University of Birmingham at TAC"*, In TAC, 2008.
7. Sekine, S. and Nobata, C., "*Sentence Extraction with Information Extraction Techniques",* Workshop on Text Summarization, 2001.
8. Jones, K. S., "*Automatic Summarizing: Factors and Directions",* Advances in automatic text summarization, pp. 1-12, 1990.
9. Nenkova, A., & McKeown, K., "*A survey of text summarization techniques*". In Mining text data, Springer, Boston, MA, pp. 43-76, 2012.
10. Zhang, Y., Chu, C. H., Ji, X., & Zha, H., "*Correlating summarization of multi-source news with k-way graph bi-clustering*", ACM SIGKDD Explorations Newsletter, 6(2), pp. 34-42, 2004.
11. Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M, "*Multi-document summarization by sentence extraction*". In Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization-Volume 4. Association for Computational Linguistics, pp. 40-48, 2000.
12. Hahn, U., & Mani, I., "*The challenges of automatic summarization*". Computer, 33(11), PP. 29-36, 2000.

13. Gong, Y., & Liu, X., "*Generic text summarization using relevance measure and latent semantic analysis*", In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, PP. 19-25, 2001.
14. Ježek, K., & Steinberger, J., "*Automatic Text Summarization (The state of the art 2007 and new challenges)*", In Proceedings of Znalosti, PP. 1-12, 2008.
15. López, M. J. M., de Buenaga Rodríguez, M., & Hidalgo, J. M. G., "*Using and evaluating user directed summaries to improve information access*", In Proceedings of the third european conference on research and advanced technology for digital libraries, LNCS 1696: Springer-Verlag, pp. 198–214, 1999.
16. Díaz, A., & Gervás, P., "*User-model based personalized summarization*", Information Processing & Management 43.6, pp. 1715-1734, 2007.
17. Adovamavicius, G., & Tuzhilin, A., "*Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*" Knowledge and Data Engineering, IEEE Transactions on 17.6, pp.734-749, 2005.
18. Das, A. S., Datar, M., Garg, A., & Rajaram, S., "*Google news personalization: scalable online collaborative filtering*", Proceedings of the 16th international conference on World Wide Web, ACM , PP. 271-280, 2007.
19. Dwivedi, S. K., & Arya, C., "*A survey of news recommendation approaches*", In ICT in Business Industry & Government (ICTBIG), International Conference on IEEE, pp. 1-6, 2016.
20. Hassel, M., & Mazdak, N., "*FarsiSum: a Persian text summarizer*", In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages Association for Computational Linguistics, pp. 82-84, 2004.
21. Saggion, H., Radev, D., Teufel, S., Lam,W., "*Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics",* In: Proceedings of COLING 2002, Taipei, Taiwan, pp. 849–855, 2002.
22. Mani, I., "*Automatic summarization*" (Vol. 3). John Benjamins Publishing, 2001.
23. Saggion, H., "*Multilingual Multidocument Summarization Tools and Evaluation"*, In: Proceedings of LREC, 2006.

24. Hovy, E., & Lin, C. Y., "*Automated text summarization and the SUMMARIST system*". In Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, Association for Computational Linguistics, PP. 197-214, 1998.
25. Litvak, M., Last, M., Friedman, M., & Kisilevich, S., "*MUSE–a multilingual sentence extractor*", Computational linguistics & applications (CLA 11), Jachranka, http://bib. dbvis. de/uploadedFiles/362. pdf (2011).
26. Mihalcea, R., "*Language independent extractive summarization",* In: AAAI, pp. 1688–1689, 2005.
27. Saggion, H., "*Experiments on semantic-based clustering for cross-document coreference",* In: Proceedings of the Third Joint International Conference on Natural Language Processing, AFNLP, AFNLP, Hyderabad, India, pp. 149–156, 2008.
28. Kabadjov, M., Atkinson, M., Steinberger, J., Steinberger, R., & Van Der Goot, E., "*Newsgist: A multilingual statistical news summarizer",* In: ECML/PKDD (3), PP. 591–594, 2010.
29. Mani, I., & Bloedorn, E, "*Multidocument summarization by graph search and matching*". In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-97), PP. 622-628, Providence, Rhode Island. AAAI, 1997.
30. Christie, F., & Khodra, M. L., "*Multi-document summarization using sentence fusion for Indonesian news articles"*. In Advanced Informatics: Concepts, Theory And Application (ICAICTA), International Conference On IEEE, PP. 1-6, 2016.
31. Gholamrezazadeh, S., Salehi, M. A., & Gholamzadeh, B., "*A comprehensive survey on text summarization systems",* In *Computer Science and its Applications, CSA'09. 2nd International Conference on* IEEE, PP. 1-6. 2009.
32. Tsoumou, E. S. L., Lai, L., Yang, S., & Varus, M. L., "*An Extractive Multi-document Summarization Technique Based on Fuzzy Logic Approach",* In Network and Information Systems for Computers (ICNISC), 2016 International Conference on IEEE PP. 346-351, 2016.
33. Fattah, M. A., & Ren, F., "*Automatic Text Summarization*", In proceedings of World Academy of Science, Engineering and Technology Volume 27, PP. 192-195, 2008.

34. Lin, C.Y., "*Training a selection function for extraction*", In Proceedings of the eighth international conference on Information and knowledge management, Kansas City, Missouri, United States. PP.55–62, 1999.
35. Kupiec, J. Pedersen, J., & Chen, F., "*A Trainable Document Summarizer*", In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Seattle, WA, PP.68-73,1995.
36. Salton, G., & Buckley, C., "*Term-weighting approaches in automatic text retrieval",* Information processing & management, *24*(5), PP. 513-523, 1988.
37. Salton, G., "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*", Addison-Wesley Publishing Company.1989.
38. Elhadad, M., "*Using Argumentation to Control Lexical Choice: A Functional Unification-Based Approach*", Ph.D. dissertation, Columbia University. 1992.
39. Ko, Y., & Seo, J., "*An effective sentence-extraction technique using contextual information and statistical approaches for text summarization",* Pattern Recognition Letters, 29(9), pp. 1366-137, 2008.
40. Wasson, M., "*Using leading text for news summaries: Evaluation results and implications for commercial summarization applications*", in Proc. 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL, 1998, pp.1364-1368.
41. AlSanie, W., Touir, A., & Mathkour, H., "*Towards an infrastructure for Arabic text summarization using rhetorical structure theory*", Master Thesis, Department of computer science. King Saud University, Riyadh, Kingdom of Saudi Arabia, 2005.
42. Kruengkrai, C., & Jaruskulchai, C., "*Generic text summarization using local and global properties of sentences",* In Web Intelligence, WI 2003. Proceedings. IEEE/WIC International Conference on IEEE, pp. 201-206, 2003.
43. Mann, W. C., & Thompson, S. A., "*Rhetorical structure theory: Toward a functional theory of text organization"*. Text-Interdisciplinary Journal for the Study of Discourse, *8*(3), pp. 243-281, 1988.
44. Marcu, D., "*Discourse trees are good indicators of importance in text*", Advances in automatic text summarization, 293, pp. 123-136, 1999.

45. Cristea, D., Postolache, O., & Pistol I., "*Summarisation through discourse structure*", In: Proceedings of the computational linguistics and intelligent text processing, 6th International conference (CICLing 2005), PP. 632–644, 2005.
46. Gonçalves, P. N., Rino L., Vieira R., "*Summarizing and referring: towards cohesive extracts",* In: DocEng '08: proceeding of the eighth ACM symposium on document engineering, PP. 253–256, 2008.
47. Baldwin, B., & Morton, T. S., "*Dynamic coreference-based summarization",* In Proceedings of the Third Conference on Empirical Methods for Natural Language Processing, PP. 1-6, 1988.
48. Azzam, S., Humphreys, K., & Gaizauskas, R., "*Using coreference chains for text summarization",* In Proceedings of the Workshop on Coreference and its Applications, Association for Computational Linguistics, PP. 77-84, 1999.
49. Barzilay R, Elhadad M., "*Using lexical chains for text summarization",* In: Advances in automatic text summarization. PP. 111–122, 1999.
50. Medelyan O., "*Computing lexical chains with graph clustering*", In: Proceedings of the ACL student research workshop, PP. 85–90, 2007.
51. Ercan G, Cicekli I, "*Lexical cohesion based topic modeling for summarization*", In: Proceedings of the 9th international conference in computational linguistics and intelligent text processing. PP. 582–592, 2008.
52. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D., & Wagstaff, K., "*Multidocumenet Summarization via Information Extraction*", Proceedings of the first International Conference on Human Language Technology Research, PP. 1-7, 2001.
53. Genest, P. E., & Lapalme, G., "*Fully Abstractive Approach to Guided Summarization*", Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 354-358, 2012.
54. Khan, A., Salim, N., & Kumar, Y. J., "*A framework for multi-document abstractive summarization based on semantic role labeling",* Applied Soft Computing, 30, PP. 737-747, 2015.
55. Barzilay, R., McKeown, K. R., & Elhadad, M., "*Information fusion in the context of multi-document summarization",* Proceedings of the 37th annual meeting of the

Association for Computational Linguistics on Computational Linguistics, (LCL' 99), Stroudsburg, PA, USA, PP. 550-557, 1999.

56. Harabagiu, S. M., & Lacatusu, F., "*Generating single and multi-document summaries with gistexter*", In Document Understanding Conferences, pp. 11-12, 2002.
57. Lammari, N., & Métais, E., "*Building and maintaining ontologies: a set of algorithms*", Data & Knowledge Engineering, 48(2), pp. 155-176, 2004.
58. Soo, V. W., & Lin, C. Y., "*Ontology-based information retrieval in a multi-agent system for digital library*", In 6th Conference on Artificial Intelligence and Applications, pp. 241-246, 2001.
59. Lee, C. S., Jian, Z. W., & Huang, L. K., "*A fuzzy ontology and its application to news summarization"*, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35(5), pp. 859-880, 2005.
60. Moratanch, N., & Chitrakala, S., "*A survey on abstractive text summarization*", In Circuit, Power and Computing Technologies (ICCPCT), 2016 International Conference on IEEE, pp. 1-7, 2016.
61. Greenbacker, C. F., "*Towards a framework for abstractive summarization of multimodal documents*", In Proceedings of the ACL Student Session. Association for Computational Linguistics, pp. 75-80, 2011.
62. Genest, P. E., & Lapalme, G., "*Framework for abstractive summarization using text-to-text generation*", In Proceedings of the Workshop on Monolingual Text-To-Text Generation, Association for Computational Linguistics, pp. 64-73, 2011.
63. Moawad, I. F., & Aref, M., "*Semantic graph reduction approach for abstractive Text Summarization*", In Computer Engineering & Systems (ICCES), Seventh International Conference on IEEE, pp. 132-138. 2012.
64. Nenkova, A., & McKeown, K., "*Automatic summarization*", Foundations and Trends® in Information Retrieval, 5(2–3), pp. 103-233, 2011.
65. Becher, M., Endres-Niggemeyer, B. and Fichtner, G., "*Scenario forms for web information seeking and summarizing in bone marrow transplantation",* Proceedings of the Conference on Multilingual Summarization and Question Answering, (SQA' 02), Stroudsburg, PA, USA, PP. 1-8, 2002.

66. Kaicker, J., Debono, V. B., Dang, W., Buckley, N., & Thabane, L., "*Assessment of the quality and variability of health information on chronic pain websites using the DISCERN instrument"*, BMC Med., 8: PP. 59-59, 2010.
67. Elhadad, N., Kan, M. Y., Klavans, J. L., & McKeown, K. R., "*Customization in a unified framework for summarizing medical literature*", Artificial intelligence in medicine, 33(2), pp. 179-198, 2005.
68. Kan, M. Y., McKeown, K. R., & Klavans, J. L., "*Applying natural language generation to indicative summarization",* Proceedings of the 8th European Workshop on Natural Language Generation, (NLC' 01), Stroudsburg, PA, USA, PP. 1-9, 2001.
69. Khelif, K., Dieng-Kuntz, R., & Barbry, P., "*An ontology-based approach to support text mining and information retrieval in the biological domain"*, J. Univ. Comput. Sci., 13: PP. 1881-1907, 2007.
70. Nenkova, A., & Bagga, A., "*Facilitating email thread access by extractive summary generation"*, Hohn Benjamins Publishing Company, 2004.
71. Newman, P. S., & Blitzer, J. C., "*Summarizing archived discussions: A beginning",* Proceedings of the 8th International Conference on Intelligent user Interfaces, Jan. 12-15, Miami, FL, USA, PP. 273-276, 2003.
72. Radev, D., Otterbacher, J., Winkel, A., & Blair-Goldensohn, S, "*NewsInEssence: summarizing online news topics",* Communications of the ACM, 48(10), pp. 95-98, 2005.
73. Mitchell, C. C., & West, M. D, "*The news formula: A concise guide to news writing and reporting"*. St. Martin's Press, 1996.
74. McKeown, K., & Radev, D. R., "*Generating summaries of multiple news articles"* In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 74-82, ACM, 1995.
75. Sebastiani, F., "*Machine learning in automated text categorization*", ACM computing surveys (CSUR), 34(1), pp. 1-47, 2002.
76. Michalski, R. S., Carbonell, J. G., and Mitchell, T. M. (Eds.), "*Machine learning: An artificial intelligence approach",* Springer Science and Business Media, 2013.
77. Mantaras,D., Lopez, R. and Plaza, E. eds, "*machine learning ECML 2000*", Springer, Vol. 1810, 2000.

78. Karamcheti, A. C, "*A Comparative study on text categorization"* (Doctoral dissertation, University of Nevada, Las Vegas) 2010.
79. Kotsiantis, S. B., Zaharakis, I., and Pintelas, P., "*Supervised machine learning: A review of classification techniques*", 2007.
80. Lu, T. T., "*Fundamental limitations of semi-supervised learning*", 2009.
81. Hu, R., "*Active learning for text classification*", 2011.
82. Ting, S. L., Ip, W. H., and Tsang, A. H., "*Is Naive Bayes a good classifier for document classification?*", International Journal of Software Engineering and Its Applications, 5(3), pp. 37-46, 2011.
83. Platt, J. C., "*Fast training of support vector machines using sequential minimal optimization*", Advances in kernel methods—support vector learning, 3, 1999.
84. Quinlan, J. R., "*Induction of decision trees. Machine learning", 1*(1), pp. 81-106, 1986.
85. Jones K. S., "*Automatic Summarising: The State of the Art*", Inf Process Manag 43(6), pp. 1449–1481, 2007.
86. Carbonell, J.G., Goldstein, J., "*The use of MMR, diversity-based reranking for reordering documents and producing summaries*", In Research and Development in Information Retrieval, pp. 335–336, 1998.
87. Radev, D.R., Jing, H., Budzikowska, M., "*Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies*", In ANLP/NAACL Workshop on Summarization. Seattle, WA, 2000.
88. Saggion, H., Gaizauskas, R., "*Multi-document summarization by cluster/profile relevance and redundancy removal*", In Proceedings of the Document Understanding Conference NIST, Boston, USA, 2004.
89. Dalianis, H., Hassel, M., de Smedt, K., Liseth, A., Lech, T. C., & Wedekind, J., "*Porting and evaluation of automatic summarization*", In Nordisk Sprogteknologi, pp. 107–121, 2004.
90. Hovy, E. & Lin, C. Y., "*Automated Text Summarization in SUMMARIST*", In I. Mani and M.T. Maybury, editors, Advances in Automatic Text Summarization, The MIT Press, pp. 81–94, 1999.

91. Saggion, H., Radev, D., Teufel, S., Lam, W., & Strassel, S. M., "*Developing infrastructure for the evaluation of single and multi-document summarization systems in a cross-lingual environment*", Ann Arbor, 1001(48), pp. 109-1092, 2002.
92. Wong, K. F., Wu, M., & Li, W., "*Extractive summarization using supervised and semi-supervised learning*", In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 985-992, 2008.
93. Qi, X., & Davison, B. D., "*Web page classification: Features and algorithms*", ACM computing surveys (CSUR), 41(2), 12, 2009.
94. Yang, Y., & Liu, X., "*A re-examination of text categorization methods*", In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval ACM, pp. 42-49, 1999.
95. Kim, S. M., Pantel, P., Duan, L., & Gaffney, S., "*Improving web page classification by label-propagation over click graphs*", In Proceedings of the 18th ACM conference on Information and knowledge management, ACM, pp. 1077-1086, 2009.
96. Dumais, S., & Chen, H., "*Hierarchical classification of Web content*", In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 256-263, 2000.
97. Cruz, I. F., Borisov, S., Marks, M. A., & Webb, T. R., "*Measuring structural similarity among web documents: preliminary results*", In Electronic Publishing, Artistic Imaging, and Digital Typography, Springer, Berlin, Heidelberg, pp. 513-524, 1998.
98. Wong, W. C., & Fu, A. W. C., "*Finding Structure and Characteristics of Web Documents for Classification*", In ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery No. s 1, pp. 96-105, 2000.
99. Agrawal, R., & Srikant, R., "On integrating catalogs", In Proceedings of the 10th international conference on World Wide Web ACM, pp. 603-612, 2001.
100. Sun, A., Lim, E. P., & Ng, W. K., "*Web classification using support vector machine*", In Proceedings of the 4th international workshop on Web information and data management ACM, pp. 96-99, 2002.
101. Riboni, D., "*Feature selection for web page classification",* pp. 1-5, 2002 na.

102. Joshi, S., Agrawal, N., Krishnapuram, R., & Negi, S., "*A bag of paths model for measuring structural similarity in Web documents*", In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp. 577-582, 2003.
103. Holden, N., & Freitas, A. A., "*Web page classification with an ant colony algorithm*". In *International Conference on Parallel Problem Solving from Nature* Springer, Berlin, Heidelberg, pp. 1092-1102, 2004.
104. Kan, M. Y., "*Web page classification without the web page*", In Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, WWW Alt. '04, New York, NY, USA ACM, pp. 262-263, 2004.
105. Selamat, A., & Omatu, S., "*Web page feature selection and classification using neural networks*", Information Sciences, 158, pp. 69-88, 2004.
106. Kan, M. Y., & Thi, H. O. N., "*Fast webpage classification using URL features*", In Proceedings of the 14th ACM international conference on Information and knowledge management, ACM, pp. 325-326, 2005.
107. Tombros, A., & Ali, Z., "*Factors affecting web page similarity*", In European Conference on Information Retrieval, Springer, Berlin, Heidelberg, pp. 487-501, 2005 .
108. Tongchim, S., Sornlertlamvanich, V., & Isahara, H., "*Classification of news web documents based on structural features*", In Advances in Natural Language Processing, Springer, Berlin, Heidelberg, pp. 153-160, 2006.
109. Chy, A. N., Seddiqui, M. H., & Das, S., "*Bangla news classification using naive Bayes classifier",* In Computer and Information Technology (ICCIT), 2013 16th International Conference on, IEEE, pp. 366-371, 2014.
110. Singh, D. A. A. G., Leavline, E. J., Priyanka, E., & Sumathi, C., "*Feature selection using rough set for improving the performance of the supervised learner*", *International Journal of Advanced Science and Technology*, pp. *87*, 1-8, 2016.
111. Muslea, I., Minton, S., & Knoblock, C., "*A hierarchical approach to wrapper induction*", In Proceedings of the third annual conference on Autonomous Agents ACM, pp. 190-197, 1999.

112. Lin, S. H., & Ho, J. M, "*Discovering informative content blocks from Web documents*", In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining ACM, pp. 588-593, 2002.
113. Knoblock, C. A., Lerman, K., Minton, S., & Muslea, I., "*Accurately and reliably extracting data from the web: A machine learning approach*", In Intelligent exploration of the web Physica, Heidelberg, pp. 275-287, 2003.
114. Reis, D. D. C., Golgher, P. B., Silva, A. S., & Laender, A., "*Automatic web news extraction using tree edit distance*", In Proceedings of the 13th international conference on World Wide Web ACM, pp. 502-511, 2004.
115. Gupta, S., Kaiser, G. E., Grimm, P., Chiang, M. F., & Starren, J., "*Automating content extraction of html documents*", World Wide Web, 8(2), pp. 179-224, 2005.
116. Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., & Etzioni, O., "*Open information extraction from the web*", In *IJCAI* Vol. 7, pp. 2670-2676, 2007.
117. Gibson, J., Wellner, B., & Lubar, S., "*Adaptive web-page content identification*", In Proceedings of the 9th annual ACM international workshop on Web information and data management ACM, pp. 105-112, 2007.
118. Ziegler, C. N., & Skubacz, M., "Content extraction from news pages using particle swarm optimization on linguistic and structural features", In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, pp. 242-249, 2007.
119. Prasad, J., & Paepcke, A., "*Coreex: content extraction from online news articles*", In Proceedings of the 17th ACM conference on Information and knowledge management ACM, pp. 1391-1392, 2008.
120. Louvan, S., "*Extracting the Main Content from HTML Documents*", Online], http://wwwis. win. tue. nl/bnaic2009/papers/bnaic2009_paper_113. Pdf, 2009.
121. Ji, X., Zeng, J., Zhang, S., & Wu, C., "*Tag tree template for Web information and schema extraction*", Expert systems with Applications, 37(12), pp. 8492-8498, 2010.
122. Guo, Y., Tang, H., Song, L., Wang, Y., & Ding, G., "*ECON: an approach to extract content from web news page*", In Web Conference (APWEB), 2010 12th International Asia-Pacific IEEE, pp. 314-320, 2010.

123. Sleiman, H. A., & Corchuelo, R., "*Tex: An efficient and effective unsupervised web information extractor*", Knowledge-Based Systems, 39, pp. 109-123, 2013.
124. Kaddu, M. R., & Kulkarni, R. B., "*To extract informative content from online web pages by using hybrid approach*", In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on IEEE, pp. 972-977, 2016.
125. Pettersson, E., Lindström, J., Jacobsson, B., & Fiebranz, R., "*HistSearch-Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text*", In HistoInformatics@ DH, pp. 25-36, 2016.
126. Chien, L. F., "*PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval*", Information processing & management, *35*(4), 501-521, 1999.
127. Turney, P.D., "*Learning to Extract Keyphrases from Text*", National Research Council, Institute for Information Technology, Technical Report ERB-1057, 1999.
128. Martínez-Fernández, J. L., García-Serrano, A., Martínez, P., & Villena, J., "*Automatic keyword extraction for news finder*", In International Workshop on Adaptive Multimedia Retrieval, Springer, Berlin, Heidelberg, pp. 99-119, 2003.
129. Wu, Y. F. B., Li, Q., Bot, R. S., & Chen, X., "*KIP: a keyphrase identification program with learning functions*", In Information Technology: Coding and Computing, Proceedings. ITCC 2004. International Conference on IEEE, Vol. 2, pp. 450-454, 2004.
130. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., & Nevill-Manning, C. G., "*KEA: Practical Automated Keyphrase Extraction*", In Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI Global, pp. 129-152, 2005.
131. Wang, J., Peng, H., & Hu, J. S., "Automatic keyphrases extraction from document using neural network", In Advances in Machine Learning and Cybernetics, Springer, Berlin, Heidelberg, pp. 633-641, 2006.
132. Lui, Y. J., Brent, R., & Calinescu, A., "*Extracting significant phrases from text*", In Advanced Information Networking and Applications Workshops, AINAW'07. 21st International Conference on IEEE, Vol. 1, pp. 361-366, 2007.
133. Wang, C., Zhang, M., Ru, L., & Ma, S., "*An automatic online news topic keyphrase extraction system*", In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01, IEEE Computer Society, pp. 214-219, 2008.

134. Xie, F., Wu, X., & Hu, X., "*Keyphrase extraction based on semantic relatedness*", In Cognitive Informatics (ICCI), 9th IEEE International Conference on IEEE, pp. 308-312, 2010.
135. Gao Y, Liu J, Ma P., "*The hot keyphrase extraction based on tf* pdf*", In Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE 10th International Conference on, pp. 1524-1528, 2011.
136. Li Z. F., Zhao X. H., Yi J, & He B., "*Improvement of KEA Based on Lexical Chain*", In Advanced Materials Research; 756, pp. 2999-3004, 2013.
137. Luo Z, Tang J, Wang T., "*Improving keyphrase extraction from web news by exploiting comments information*", In Asia-Pacific Web Conference, pp. 140-150, 2013.
138. Li, Z. & He, B., "*Adding Lexical Chain to Keyphrase Extraction*", In Web Information System and Application Conference (WISA), pp. 254-257, 2014.
139. Hsu, H. M., Chang, R. I., Chang, Y. J., Lin, S. Y., Wang, Y. J., & Ho, J. M., "*Subject-Keyphrase Extraction Based on Definition-Use Chain*", In Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE/WIC/ACM International Conference on IEEE, Vol. 3, pp. 199-202, 2015.
140. Duwairi R, Hedaya M., "*Automatic keyphrase extraction for Arabic news documents based on KEA system*", Journal of Intelligent & Fuzzy Systems, 30(4), pp. 2101-10, 2016.
141. Grineva, M., Grinev, M., & Lizorkin, D., "*Extracting key terms from noisy and multitheme documents*", In Proceedings of the 18th international conference on World wide web ACM, pp. 661-670, 2009.
142. Ercan G, Cicekli I., "*Using lexical chains for keyword extraction*", Information Processing & Management, 43(6), pp. 1705-14, 2007.
143. Sundheim, B. M., "*Overview of the fourth message understanding evaluation and conference*", In Proceedings of the 4th conference on Message understanding, Association for Computational Linguistics, pp. 3-21, 1992.
144. Chen, H. H., & Lin, C. J., "*A multilingual news summarizer*", In Proceedings of the 18th conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pp. 159-165, 2000.

145. White, M., Korelsky, T., Cardie, C., Ng, V., Pierce, D. & Wagstaff, K., "*Multidocument summarization via information extraction*", Proceedings of the 1st International Conference on Human Language Technology Research, (LTR' 01), pp. 1-7, 2001.
146. McKeown, K. R., Hatzivassiloglou, V., Barzilay, R., Schiffman, B., Evans, D., & Teufel, S., "*Columbia multi-document summarization: Approach and evaluation*", in Proc. DUC Conf. Text Summarization, 2001.
147. McKeown, K. R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J. L., Nenkova, A., & Sigelman, S., "*Tracking and summarizing news on a daily basis with Columbia's Newsblaste*r", In Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., pp. 280-285, 2002.
148. Daniel, N., Radev, D., & Allison, T., "*Sub-event based multi-document summarization*", Proceedings of the HLT-NAACL on Text Summarization Workshop, (TSW' 03), ACM, pp. 9-16, 2003.
149. Lee, C. S., Chen, Y. J., & Jian, Z. W., "*Ontology-based fuzzy event extraction agent for chinese e-news summarization*", Expert Systems With Applications, vol. 25, no. 3, pp. 431–447, 2003
150. D'Avanzo, E., & Magnini, B., "*A keyphrase-based approach to summarization: the lake system at duc-2005*", In Proceedings of DUC, 2005.
151. Svore, K., Vanderwende, L., & Burges, C., "*Enhancing single-document summarization by combining RankNet and third-party sources*", In Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), 2007.
152. Litvak, M., & Last, M., "*Graph-based keyword extraction for single-document summarization*", In Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, Association for Computational Linguistics, pp. 17-24, 2008.
153. Li, L., Wang, D., Shen, C., & Li, T., "*Ontology enriched multi-document summarization in disaster management*", Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Jul. 19-23, Geneva, Switzerland, pp: 819-820, 2010.

154. Al-Hashemi, R., "*Text Summarization Extraction System (TSES) Using Extracted Keywords*", Int. Arab J. e-Technol., 1(4), pp. 164-168, 2010.
155. Litvak, M., Last, M., & Friedman, M., "*A new approach to improving multilingual summarization using a genetic algorithm*", In Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics, pp. 927-936, 2010.
156. El-Haj, M., Kruschwitz, U., & Fox, C., "*Exploring clustering for multi-document arabic summarisation*", Asia Information Retrieval Symposium. Springer, Berlin, Heidelberg, 2011.
157. Galanis, D., Lampouras, G., and Androutsopoulos, I., "*Extractive multi-document summarization with integer linear programming and support vector regression*", Proceedings of COLING, pp. 911-926, 2012.
158. Li, C., Qian, X., & Liu, Y., "*Using supervised bigram-based ilp for extractive summarization*", In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vol. 1, pp. 1004-1013, 2013.
159. Yan, S., & Wan, X., "*SRRank: leveraging semantic roles for extractive multi-document summarization*", IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), 22(12), pp. 2048-2058, 2014.
160. Liu, M., Wang, L., & Nie, L., "*Weibo-oriented chinese news summarization via multi-feature combination*", In Natural Language Processing and Chinese Computing, Springer, Cham, pp. 581-589, 2015.
161. Cao, Z., Wei, F., Dong, L., Li, S., & Zhou, M., "*Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization*", In AAAI, pp. 2153-2159, 2015.
162. Liu, S. H., Chen, K. Y., Hsieh, Y. L., Chen, B., Wang, H. M., Yen, H. C., & Hsu, W. L., "*Exploring Word Mover's Distance and Semantic-Aware Embedding Techniques for Extractive Broadcast News Summarization*", In INTERSPEECH, pp. 670-674, 2016.
163. Demirci, F., Karabudak, E., & İlgen, B., "*Multi-document summarization for Turkish news*", In Artificial Intelligence and Data Processing Symposium (IDAP), International, IEEE, pp. 1-5, 2017.

164. Radev, R., Blair-goldensohn, S, Zhang, Z., "*Experiments in single and Multi-documents Summarization using MEAD*", In First Document Understanding Conference. New Orleans, LA, 2001.
165. Vanderwende, L., Banko, M., & Menezes, A., "Event-centric summary generation", Working notes of DUC, pp. 127-132, 2004**.**
166. Fuentes, M., Rodr´ıguez, H., Ferres, D., "*FEMsum at DUC 2007*", In the Document Understanding Workshop (presented at the HLT/NAACL), Rochester, New York USA, 2007.
167. An, A., Huang, Y., Huang, X., & Cercone, N., "*Feature selection with rough sets for web page classification*", In *Transactions on rough sets II*, Springer, Berlin, Heidelberg, pp. 1-13, 2004.
168. Choi, B., & Yao, Z., "*Web page classification*", In Foundations and Advances in Data Mining, Springer, Berlin, Heidelberg, pp. 221-274, 2005.
169. Pechenizkiy, M., "*The impact of feature extraction on the performance of a classifier: kNN, Naïve Bayes and C4.5",* In Conference of the Canadian Society for Computational Studies of Intelligence. Springer, Berlin, Heidelberg, pp. 268-279, 2005.
170. Wu, X., Wu, G. Q., Xie, F., Zhu, Z., & Hu, X. G., "*News filtering and summarization on the web*", *IEEE Intelligent Systems*, *25*(5), pp. 68-76, 2010.
171. Ramdass, D., & Seshasai, S., "*Document classification for newspaper articles*", Document classification for newspaper articles, 2009.
172. Lewis, D. D., "*Naive (bayes) at forty: The independence assumption in information retrieval*", Machine Learning: ECML-98, pp. 4–15, 1998
173. Holmes, G., Donkin, A., & Witten, I. H., "*Weka: A machine learning workbench*", In Intelligent Information Systems, 1994. Proceedings of the 1994 Second Australian and New Zealand Conference on IEEE, pp. 357-361, 1994.
174. Patil, A. S., & Pawar, B. V., "*Automated classification of web sites using Naive Bayesian algorithm*", In Proceedings of the international multiconference of engineers and computer scientists, Vol. 1, pp. 519-523, 2012.
175. Dwivedi, S. K., & Arya, C., "*Automatic Text Classification in Information retrieval: A Survey*", In Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies, ACM, pp. 131, 2016.

176. http://text-processing.com/demo/tokenize/
177. Chang, C. H., & Lui, S. C., "*IEPAD: information extraction based on pattern discovery*", In Proceedings of the 10th international conference on World Wide Web, ACM, pp. 681-688, 2001.
178. Raggett, D., "*Clean up your Web pages with HP's HTML tidy*", Computer networks and ISDN systems, 30(1-7), pp. 730-732, 1998.
179. Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., & Nevill-Manning, C. G., "*Domain-specific keyphrase extraction*", In 16th International joint conference on artificial intelligence (IJCAI 99), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Vol. 2, pp. 668-673, 1999 .
180. Miller, G., "WordNet: An electronic lexical database", MIT press, 1998.
181. http://nlp.stanford.edu/software/tagger.shtml
182. Morris J, Hirst G., "*Lexical cohesion computed by thesaural relations as an indicator of the structure of text*", Computational linguistics, 17(1), pp. 21-48, 1991.
183. Steffen R., "*Lexical chain Annotation Guidelines*", 2012.
184. Silber, H. G., & McCoy, K. F., "*Efficiently computed lexical chains as an intermediate representation for automatic text summarization*", Computational Linguistics, 28(4), pp. 487-496, 2002.
185. Barzilay, R., "*Lexical chains for summarization",* (Doctoral dissertation, Ben-Gurion University of the Negev), 1997.
186. Fragos, K., Maistros, Y., & Skourlas, C., "*Word sense disambiguation using wordnet relations*", In First Balkan Conference in Informatics, Thessaloniki, 2003.
187. Wei, T., Lu, Y., Chang, H., Zhou, Q., & Bao, X., "*A semantic approach for text clustering using WordNet and lexical chains*", Expert Systems with Applications, 42(4), pp. 2264-2275, 2015.
188. Sussna, M., "*Word sense disambiguation for free-text indexing using a massive semantic network*", In Proceedings of the second international conference on Information and knowledge management, ACM, pp. 67-74, 1993.
189. Siblini, R., & Kosseim, L., "*Using a weighted semantic network for lexical semantic relatedness*", In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP, pp. 610-618, 2013.

190. Suo, H., Liu, Y., & Cao, S., "*A keyword selection method based on lexical chains*", Journal of Chinese Information Processing, 20(6), pp. 25-30, 2006.
191. Qazvinian, V., Radev, D. R., & Özgür, A., "*Citation summarization through keyphrase extraction*", In Proceedings of the 23rd international conference on computational linguistics, Association for Computational Linguistics, pp. 895-903, 2010.
192. McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E., "*Towards multidocument summarization by reformulation: Progress and prospects*", In Proceedings of AAAI-99, 1999.
193. Nenkova, A., & Passonneau, R., "*Evaluating content selection in summarization: The pyramid method*", In Proceedings of the human language technology conference of the north american chapter of the association for computational linguistics: Hlt-naacl, 2004.
194. Shen, C., & Li, T., "*Multi-document summarization via the minimum dominating set*", In Proceedings of the 23rd International Conference on Computational Linguistics, Association for Computational Linguistics, pp. 984-992, 2010.
195. Lin, C. Y., "*Rouge: A package for automatic evaluation of summaries*", Text Summarization Branches Out, 2004.
196. Chin-Yew, L., & Och, F. J., "*Looking for a few Good Metrics: ROUGE and its Evaluation*", National institute of Informatics, Working notes of NTCIR-4, Tokyo, 2004.
197. Lloret, E., & Palomar, M., "*Challenging issues of automatic summarization: relevance detection and quality-based evaluation*", Informatica, 34(1), 2010.
198. Alguliev, R. M., Aliguliyev, R. M., & Mehdiyev, C. A., "*pSum-Sade: a modified p-median problem and self-adaptive differential evolution algorithm for text summarization*", Applied Computational Intelligence and Soft Computing, pp. 11, 2011.

Appendix

APPENDIX-I: LIST OF PUBLICATIONS

1. Arya, Chandrakala, and Dwivedi, Sanjay K., "Content extraction from news web pages using tag tree," International Journal of Autonomic Computing, 3(1), pp. 34-51, **Inderscience (ACM Digital Library),** 2018.

2. Arya, Chandrakala, and Dwivedi, Sanjay K., "Keyphrase Extraction of News Web Pages," International journal of Education and Management Engineering (IJME), vol. 8 (1), pp. 48-58, **MECS**, 2018.

3. Arya Chandrakala, and Dwivedi Sanjay K, "Keyphrase-based Multi-Document News Web Pages Summarization," in the Computer Journal **SCI, IF: 0.792.** (Communicated).

4. Dwivedi, Sanjay K., and Arya, Chandrakala, "A survey of news recommendation approaches," In ICT in Business Industry & Government (ICTBIG), International Conference on, pp. 1-6, **IEEE**, 2016.

5. Arya, Chandrakala and Dwivedi, Sanjay K., "News web page classification using url content and structure attributes," Next Generation Computing Technologies (NGCT), 2nd International Conference on, pp. 317-322, **IEEE**, 2016.

6. Dwivedi, Sanjay K., and Arya, Chandrakala, "Automatic Text Classification in Information retrieval: A Survey," In Proceedings of the 2nd International Conference on Information and Communication Technology for Competitive Strategies, pp. 131, **ACM Digital Library,** 2016.

7. Arya, Chandrakala, and Dwivedi, Sanjay K., "Supervised Word Sense Disambiguation in Information Retrieval: A Short Survey," published in 3rd Lucknow Science Congress & National Conference on Science for Society an Interdisciplinary Approach, 2015.

APPENDIX-II: LIST OF ABBREVIATION

S. No.	Title	Description
1	TC	Text Classification
2	WEKA	Waikato Environment for Knowledge Analysis
3	TP	True Positive
4	TN	True Negative
5	FP	False Positive
6	FN	False Negative
7	ROUGE	Recall-Oriented Understudy for Gisting Evaluation
8	DUC	Document Understanding Conference
9	TAC	Text Analysis Conference
10	POS	Part Of Speech
11	K-NN	K-Nearest Neighbour
12	SVM	Support Vector Machine
13	HTML	Hypertext Markup Language
14	NLP	Natural Language Processing
15	LSA	Latent Semantic Analysis
16	SVD	Singular Value Decomposition
17	RSG	Rich Semantic Graph
18	PDA	Personal Digital Assistants
19	URL	Uniform Resource Locator
20	IR	Information Retrieval
21	SMO	Sequential Minimal Optimization

22	TREC	Text Retrieval Conference
23	DOM	Document Object Model
24	TF-IDF	Term Frequency- Inverse Document Frequency
25	LC	Lexical Chain
26	KESR	Key-phrase Extraction based on Semantic Relations
27	KELC	Key-phrase extraction based on lexical chains
28	CART	Classification and Regression Tree
29	MDS	Multi-Document Summarization System
30	TT	Tag Tree
31	SP	Sentence Position
32	SL	Sentence Length

APPENDIX-III: ATTRIBUTES COLLECTION FROM NEWS WEBSITES

Second Level Domain	First-Level Catalog	News Center	Article Source	Author	Related News	Related Subject	Related Link	Sum Up No. of Times Term News Occur in HTML Page	Class
economy	news	march WPI inflation shrinks to 2.33 per versus 2.06 per in february	times of india	PTI	rupee convertibility needed to put india on top	inflation	foreign companies see turnaround in India says CII director general	1	news
technology	news	facebok ceo Mark Zuckerberg to HT Internet org and net neutrality can coexist	hindustan times	fp editor	backing net neutrality, clearity	internet	battel for open internet	1	news
sports	news	yuvraj, agarwal sizzle as daredevils end losing streak	times of india	Amit Karmarkar	IPL 2015 delhi daredevils beat king 11 punjab	IPL cricket	Indian premium league 2015	3	news
sports	news	saina nehwal regains world no 1 badminton ranking	hindustan times	PTI	Badminton no touranemt safe from fixing says Danish player Vittinghuse	Badminton	Airlines refuses weapons Indian shooters left in the lurch	1	news
india news	news	aap is turned into a khap expelled leaders prashant bhushan, yogendra yadav target Arvind kejriwal	ndtv	Deepshikha Gosh	both leaders hit back with scathering public replies	politics	will introduce system of Award, punishments, Arvind kejriwal tells beurocrats	1	news
sports	tennis news	Rafeal Nadal targets further improvements In barcelona	ndtv	Agence france press	Novak Djokovic relaxed over hot streaks but needs a rest	tennis	beatan in straight sets by world number one Novak Djokovic	2	news
market	news	Daiichi exit from sun pharma hits rupee stock	profit ndtv	thomson reuters	sun pharma slumps over 10 per as Daiichi exits	stock market	sensex languishes in red banking metal stock outperform	1	news
economy	news	RBI rupees management undermines hedging push	profit ndtv	reuters	unconventional monetory policies to	RBI	scope for further reduction in rates Aditya	2	news

					have negative affect Jaitly		puri		
budget	news	PPF rates unchaged but sukanya samriddhi scheme to earn more	profit ndtv	PTI	budget 2015 transport allownce exemption double	budget schemes	service tax impact air travel becomes expensive	1	news
gadgets	news	consumers withdraw lawsuits against google over Android app limits	profit ndtv	thomson reuters	EU charges google in internet search anti trust case	google search engine	sumsung electonics mobile cheif pay more than doubles in 2014	1	news
gadgets	news	Amazon launces button for instant product ordering	profit ndtv	thomson reuters	snapdeal buys freecharge in mobile transaction push	online retailers	Apple in talk to launch online TV service	2	news
careers	news	Job opportunities to grow fastest in india CFO survey	profit ndtv	PTI	Asian shares at fresh seven year high	job opportunities	Job satisfaction not necessarily equlas job loyalty	1	news
fitfoodie	yahoo insider	taste delicious racipes at saffola fit foodie	times of india	vikas khanna	fitfoodie panel	food	recipes	0	not news
india	news	taskmaster modi shows softer side to ministers	times of india	TNN	civil servent chill out PM Modi	political	political intervention in beaurocracy in necessary	3	news
sports	news	Chris smalling extends manchester united deal untill 2019	ndtv	Agence france presse	Manchester united destroy manchester city in derby	english premier league	Fabregas desperate to end wait for EPL title	2	news
Tyre knowledge	campign	know your wheels	bridgestone	Na	Basic structure of radical tyre	corporate tyre campaign	how to read tyre sidewall	0	not news
business	india business	India to become fifth largest market in the world in infrastructure projects	times of india	Nauzer Bharucha	Health infrastructure in draft development plan woefully inadequate	market	royal institution of charted surveyors	2	news
Aditya birla customer service private limited	campign	the smarter, quicker and easier way of doing SIP	myuniverse	na	Start a ZIPSIP	Aditya Birla money	link all your accounts	0	not news
zsys	campign	office lucknow	regus	NA	quick quote	landing	privacy policy	0	not news

houshold	ads	found 952 houshold items adds in lucknow	askme	NA	houshold items adds in delhi	houshold adds	used cars ads in delhi	0	not news
science	news	runway galaxies found speeding at nearly 10 milion kmph	times of india	subodh verma	ALMA peers inside starburst galaxies for the first time	Astronomy	black hole much bigger than sun found	1	news
world	news	Indian American Vivek murthy is US youngest surgeon general	times of india	Chidan and Rajghatta	Barak Obama declares cyberattacks a national emergency	achivement for india	US blame Dow flash crash on British indian trader seeks his extradiction	2	news
forum	food	chettinad food in delhi	ndtv	kriti	Hola best Maxican food in delhi	food community	recipes	0	not news
company registration	pramoted links	how to open a private limited company	Indiafills	NA	starting a business	private limited company	learning centre	0	not news
opinion	article	will real IP policy stand up	indian express	ShamnadBasheer	redraw the calender	columns	new drug era	1	news
opinion	article	the real Un question	indian express	Richard Gowan	money well spent	columns	4 lesser ways to save tax	4	news
India	article	Assam in Govt OKs India Bangladesh border swap deal	indian express	Pradeep kaushal	Aap Mujh Pe Bharosa Rakhiye PM Narendra Modi Told Sheikh Hasina	politics	Assam inclusion paves the way for boundary pact	1	news
product	compaign	sony 1000 mAh power bank	askme bazar	NA	entertainment	online shopping	terms and policies	0	not news
index	compaign	IPL 2015	India today	NA	IPL news	IPL 8	cricket section	0	not news
app data	compaign	polo hunter women watch	askme bazar	NA	our categories	online shopping	grocery	0	not news
togertheronline	ads	gift her the power of Internet this mothers day	hwgo	NA	explore the internet	together online success stories	helping women get online	0	not news

dsc	show	Dr subhash chandra show	znews	NA	Dr subhas chandra show what is the purpose of your life	tv show	about the show	0	not news
astermedcity	ads	aster medcity dedication to the nation	healthcare destination in south Asia	NA	multispeciality hospital	hospital advertisement	the concept	0	not news
travelplus	calender	indiatoday travel plus	indiatoday	NA	subscribe now	advertisement	advertise with us	0	not news
woman	shop	sarees for you	paytm	NA	mobile recharge	advertisement	sports and health	0	not news
safety shose	special offer	special brand special price	tolexo	NA	view all categories	advertisement	office supplies	0	not news
sports	football	Manchester United launch scramble to usurp Chelsea	times of india	NA	manU plays a new india plan	football match	united launch scramble to usurp chelsea	1	news
business	india business	Indias first IBU to be operational in 2 months at GIFT	times of india	Kalpesh Damor	Modi Gods gift Swaraja national asset Venkaiah	finance	Customized Supplementation vs Buying Over the Counter	2	news
business	international business	Tencent leads Rs 570cr round in Practo	times of india	NA	Practo raises dollar 90 million in fresh funding	Tencent	US oil gets cheaper on the back of innovation	3	news
tech	tech news	Texting helps Kolhapur farmers stay updated on agriculture information	times of india	Samrat phadnis	tenanat farmers still await loan eligibilty card	tech for farmers	Union agriculture ministrys mkisan web portal	4	news
tech	tech news	Foxconn signs MoU with Maharashtra government	times of india	Clara lewis	foxconn top executives meets startups in bengluru	foxconn Mou	Foxconn	4	news
sports	cricket news	Ashes World champions Australia resemble Dads Army	hindustan times	reuters	goodbye clark Aussie skipper has bittersweet end of career	cricket	England win 4 test to regain Ashes clark announces retirement	1	news
sports	football	Defending EPL champions Chelsea held Man United	hindustan times	AFP	man united get EPL season underway with	football match	Liverpool wil be better this season says manager	1	news

		open with win			win over tottenham		Rodgers		
India	india news	PM Modi to visit poll bound Bihar address rally in Gaya today	hindustan times	HT correspondent	Something wrong in Nitish political DNA Modi at Muzafferpur Rally	politics	Nitish urge PM Modi to take back comment on political DNA	3	news
India	india news	Indian Army on infiltration alert steps up LoC deployment	hindustan times	Rahul singh	two brothers held in Kashmir for sheltering Pak militant Naveed	Udhampur attack	JK police say Naveed is not a native of the state	1	news
India	india news	CBI registers five more FIRs in Vyapam scam	hindustan times	HT correspondent	we are first victims vyapam kingpin Jagdish sagar kin	vyapam	CBI files five more FIR	1	news
India	india news	Parliament Will Function If Rahul Gandhi Dares Sushma Swaraj	ndtv	Amit Chaturvedi	PM Narendra Modi addresses parivartan rally in Bihar gaya	politics	suspended lawmakers to be back in parliament today	2	news
sports	sports	India in Sri Lanka Murali Vijay Ruled Out of First Test Due to Hamstring Injury	ndtv	reuters	India set for sangkara fairwell party	cricket	Lalit Modi reveals plans to overthrow cricket world cup establishment	6	news
sports	tennis news	Rohan Bopanna Florin Mergea Bow Out of Citi Open Semis	ndtv	NA	Bopanna Margea knocked out in wimbeldon semis	tennis	Bopanna and Mergea enter wasignton quarter final	6	news
business	news	Sensex Ends 300 Points Off Days High Nifty Settles Below 8550	ndtv	Abhishek vasudev	Nifty heads for highest close in nearly three weeks	market	China says property market to continue recovering in H2	2	news
tech	news	Sundar Pichai is spicy rasam in Googles fuzzy Alphabet soup	hindustan times	Narayanan Madhavan	India born sundar Pichai appointed the new CEO of google	google	man of the moment the importance of being sundar pichai	1	news
tech	news	Foxconns focus shift India preference cause concerns in China	hindustan times	PTI	He shy bookish type Schoolmates talk about Sundar Pichai	foxconn	Xiaomi	2	news
real state	news	Yamuna Authority is	hindusta	Jeevan	NSCN IM muivah	property	why was noida authority	1	news

		offering 900 plots in Sector 22D A plot can now be yours for just Rs 17 lakh	n times	prakash sharma	under pressure to disclose Naga deal details		in a rush to approve housing project after getting a case dismissed in court		
business	news	Sensex gains 103 points in early trade driven by value buying	hindustan times	PTI	sensex slumps by 134 points Lupin Bajaj disappointing earnings	stock market	GST bill to be tabeled in RS today as cong steps up pressure	1	news
business	news	Rupee dips by 29 paise against dollar after yuan devaluation	hindustan times	PTI	China devaluates yuan	rupee	rupee against dollar	1	news
national	news	The dangerous lure of Western Ghats	The Hindu	Mohit m rao	AIUTUC critisises centre policies	national news	mysuru Dasara to be a low key affair this year	2	news
national	news	PM hails Mulayams support to end logjam in Parliament	The Hindu	PTI	parliament logjam continues	national politics	we want parliament to go on mulayam tells speaker	3	news
business	industry	Xiaomi ties up with Foxconn	The Hindu	Santosh Patnaik	Sri Lanka softens stance on China backed port project	economy	fresh low brewing in bay not much gains for penninsula	2	news
business	economy	Banks told to be sensitive to needs of MSMEs	The Hindu	Shriram Lakshman	sensex down 62 points	economy	oil sinks to a new low	2	news
sports	other sport	World Championship Kashyap Prannoy off to winning starts	The Hindu	IANS	Vijayalakshmi stays ahead in Asian women	Badminton	Bolt vs Gatlin showdown in world championship	2	news
sports	cricket news	India eyes third spot in ICC Test rankings	The Hindu	NA	injured Murali Vijay ruled out of first test	cricket	crisis hit south Africa faces rampaging Aussies	2	news
sports	hockey	Gurbaj Singh suspended for 9 months	The Hindu	PTI	Kalinga Lancers retain three indian players for HILL	Hockey India	Ranchi rays retains its best players for HILL	3	news
sports	tennis news	John Isner rallies to reach Atlanta Open final	The Hindu	AP	Venus Williams ousted in 1st round at rogers cup	tennis	serena maintains lead in WTA ranking	2	news

sci tech	health	Hyderabad scientists synthesise novel low calorie fats	The Hindu	Y Mallikarjun	spread of drug resistant malaria parasites looms large	health	brain work contributes to physical fatigue	2	news
sci tech	science	NASA drones to explore Moon and Mars	The Hindu	PTI	Rasian astronauts set for spacewalk on ISS	science	Global satellite to be named after Abdul Kalam	2	news
sci tech	energy and environment	As oceans get warmer fish are diving deeper	The Hindu	IANS	beached blue whale dies in Maharastra	environment	Indian deltas are sinking	2	news
sci tech	news	Smart villages to boost rural economy	The Hindu	Mohamed Iqbal	street food fusion festival begins tomorrow	agriculture	In search of a new route	2	news
article	India	Aadhaar card will be optional for availing social benefit schemes Supreme Court	Indian express	Utkarsh Anand	two years ago UPA discussed an offer from Dawood to return home	Aadhar card	Linking to banks continues as no order yet from centre says district collector	5	news
article	India	Nitish targets PM Modi with Shabd Wapsi campaign	Indian express	Santosh Singh	Bombay HC grants anticipatory bail to teesta setalved husband Javed Anand	politics	social harmony is in the DNA of NDA Rajnath singh	7	news
article	India	PM Narendra Modi to address UN development summit to visit San Francisco	Indian express	PTI	California plans Madison square garden like reception for PM Modi	International news	Naga pact just a formula CM T R Zeliang	5	news
sports	football	New signings settled quickly Manchester United eyeing trophies Michael Carrick	Indian express	reuters	Manchester United settling to life under Louis van gaal	football match	Manchester united shaky despite falcao boost	5	news
sports	football	Barcelona eye six trophy haul in calendar year says Andres Iniesta	Indian express	reuters	UEFA champions league final FC bercelona complete treble with 3 to 1 win over Juventus	UEFA super cup	Messi wins FIFA world player of year award	5	news

sports	tennis news	Rogers Cup Gael Monfils Gilles Simon through to second round after easy wins	Indian express	reuters	no slip ups for nafael Nadal on clay at Monte carlo masters	Rogers cup	Roger federer completes swiss comeback	5	news
sports	tennis news	Maria Sharapova pulls out of WTA event in Toronto	Indian express	reuters	Maria sharapova sets up semis date with serena williams	WTA event	Rafeal Nadal out of paris masters	5	news
article	India	Pilot meets President with sack Raje plea	indian express	express news service	ceasefire voilation pakistani troops resort to mortar shelling in poonch district	politics	Rajasthan HC says Santhara illegal jain saint wants PM Modi to move sc	6	news
article	India	Missing Hyderabad trekkers found safe in Karnataka forest	indian express	Santosh kumar	judges in HC scams as for guts I m knowing to Maya and Mulayam CJ	missing youth	Gujrat pulls books with anti Hindu Ambedkar remarks	8	news
world	Asia	Japan restarts nuclear reactor after break due to Fukushima	indian express	AP	retired US generals admirals back Iran deal say no option of rejecting it	Japan nuclear reacter	Hillary clinton finally relents agrees to give up possession of private email server	6	news
world	europe	Possible missile parts found at MH17 crash site Dutch prosecutors	indian express	AP	MH 17 wreckage to be ressambled in Netherlands for investigation	MH 17 crase	Malaysia airlines flight MH17 found with oxygen mask Dutch minister	6	news
technology	tech news technology	Did Google learn the Alphabet to get Pichai as CEO	Indian express	Tech Desk	PM Modi Satya Nadella and Tim cook congratulate Sunder Pichai	google	Sunder Pichai will always be focused on innovation says Larry page	8	news
technology	tech review	Phicomm Energy 653 Express Review A good choice at Rs 4999	Indian express	Debashis sarkar	MakeinIndia Xiaomi launches readmi 2 prime which is made in Andhra Pradesh	Xiaomi cell phone	Karbonn too eyes online only model will look to procure locally for make in India	8	news
world	neighbour	Curfew imposed army deployed after Nepal protest	Indian express	Yubaraj ghimire	syria rebels pro regime forces agree 48 hour truce monitor	curfew in Nepal	Five dead in Taliban suicide blast on kabul airport road	7	news

world	neighbour	Bangladesh police launches hunt to nab killers of blogger	Indian express	PTI	Bagladesh police asks secular bloggers not to cross limits	Bangladesh news	FBI to assist Bangladesh in blogger murder probe	5	news
sports	cricket news	Ashes 2015 Drop from squad Brad Haddin leaves for home early	indian express	AFP	criticism of families on tour makes Australiya players see red	Ashes series	Steve Smith lacking experience and support has big shoes to fill	4	news
national	news	PMs sudden UAE trip takes many by surprise	The Hindu	Suhasini Haider	Important visit to a long time allay	Modi UAE tour	general consensus on GST bill likely	5	news
national	news	Tough task ahead for Telangana	The Hindu	B Chandrashaker	Congress road blockade programme today	Telangana	celebrate Hyderabad Liberation day	2	news
International	news	Police chief in US regained control of Ferguson protests	The Hindu	AP	Coal mine explosion kills 10 in China	Ferguson protests	Avoid putting Phone number on facebook	4	news
business	news	Premji Nadar among Forbes 20 richest people in tech	The Hindu	PTI	Make hallmarking mandatory	forbes list	we aim to attract investments that can create lots of jobs	2	news
India	news	Furious Sonia Gandhi Joins Congress Protests in Well of Lok Sabha	ndtv	Deepshikha Gosh	Inside parliament fencing to keep troublemakers at bay	parliament news	let parliament function India Inc pleads in petition sighned by over 15000	2	news
football	news	Cristiano Ronaldo Sergio Ramos Join Real Madrid Practice	ndtv	Indo Asian news service	Pedro Helps Barcelona Lift UEFA Super Cup in Nine Goal Classic	Football real madrid	Sergio Ramos not leaving real madrid	8	news
business	india business	Vistara announces premium economy ticket to every business class passenger till September 30	times of india	Mihir Mishra	Vedanta slaps claims notice on govt	economy	5 shops fly out of city airport over 6 months	5	news
business	international business	China policy lenders to issue 15 billion yuan in bonds offshore	times of india	NA	Chinas move to cut currency reverberates in global economy	china economy	China induces a sharp devaluation of the yuan	3	news

market	stock news	Sensex ends 354 pts down as China GST weigh gold reclaims 26K banks bleed metals melt IT shines	times of india	economic times	the best six mid cup value stocks to buy	stock market	top six reasons why sensex slipped 551 points breached below 28 k	14	news
maraket	stock news	Yuan may fall by another 10 per from recent peak Fed unlikely to raise rates in September by Marc Faber	times of india	economic times	top 10 wealth creating ideas by experts with a holding period of 1 to 5 years	stock market	some of most influential Dalal Street investors youve never heard about	15	news
International	news	Pakistan NSA Sartaj Aziz to visit India for talks on August 23	hindustan times	HT correspondent	Pakistan readies India terror dossier for NSA meet next month	Pakistan security advisor	Modi sharif meeting all that happened behind the scenes of ufa	2	news
India	news	Govt firm on rolling out GST from next fiscal, says Javadekar	hindustan times	HT correspondent	Delay in passing GST bill will hit investments and jobs	GST bill	banned santhara ritual is not similar to suicide say jains	1	news
india	news	Pakistan national with HuJI links held in Hyderabad	hindustan times	PTI	Envoy Abdul Basit says Pakistan wont abandon kasmiris	terririst orgnisation	Modi should fulfil orop promise Rahul at jantar mantar	1	news
sports	news	Sindhus bid for 3rd straight bronze at World Championships ends	hindustan times	PTI	Gutta ponnapaa lose at world championships miss on bronze	Badminton	Delhi metro proposes fare hike and five flab	1	news
sports	news	Ajinkya Rahanes eight breaks Test fielding record	times of india	TNN	steven smith named Australias new test captain	cricket	my shot selection is much better now shikhar dawan	6	news
market	news	FIIs may pull out of India as weak Rupee may trigger sell-off on the Street	times of india	Rajesh	gold to rise on the back of weakenning rupee	stock market	Govt clears FDI proposals worth 10379crore	12	news
onlinetermlanding	onlinesales	Realiance online term	realince life	NA	Adequate cover	realiance advertisement	know your premium in seconds	0	not news
mobile applist	app	times of india for smartphones and tablets	times of india	NA	TOI for IPAD	mobile app	TOI for windows	0	not news

I.J. Education and Management Engineering, **2018, 1, 48-58**
Published Online January 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.01.06

Available online at http://www.mecs-press.net/ijeme

Keyphrase Extraction of News Web Pages

Chandrakala Arya[a*], Sanjay k. Dwivedi[b]

[a] *Research Scholar, Department of Computer Science, B.B. Ambedkar University, Lucknow-226025,India*
[b]*Professor, Department of Computer Science, B. B. Ambedkar Unversity, Lucknow-226025, India*

Received: 28 Februayr 2017; Accepted: 11 September 2017; Published: 08 January 2018

Abstract

Keyphrase extraction from news web pages is an important task for news documents retrieval and summarization. Keyphrases are like index terms that enclose the important information about document content. Keyphrases actually offer concise and precise description of document content. Key phrases are considered as a single word or a combination of more than one word that represent the important concepts in a text documents. The aim of this paper is to develop and evaluate an automatic keyphrases extraction approach for news web pages. Our approach identifies the candidate keyphrases from documents and chooses those candidate keyphrase having highest weight score. Weight formula combines the feature set that includes TF*IDF, phrase disatnce in documents and lexical chain that is based on WordNet to represent semantic relations between words. The experimental results show that the performance of our approach is better than the contemporary approaches today.

Index Terms: Keyphrase extraction, Lexical chain, Web News, TF*IDF, WordNet.

1. Introduction

Under the growth of worldwide networking through the internet, the news consumption pattern moved from the traditional physical newspapers to online news aggregate system. As thousands of web news is posted on the internet every day, it is difficult to retrieve and summarize the relevant document effectively. So keyphrase extraction technique is used to provide the main contents of a given web page. It is useful in many areas like summarization, automatic indexing, topic search and clustering [7]. Keyphrase extraction is one of the most important tasks in news web pages. Readers make benefit from keyphrase because they can judge more quickly whether the news web page is worth reading.

* Corresponding author.
E-mail address: arya.chandrakala@gmail.com, skd200@yahoo.com

Keyphrases provide a concise description of document content. We treat a document as a set of phrases; any phrase in a new document can be extracted as a keyphrase. Keyphrase can be defined as a phrase of one or more words that denote the main concept of the document. Phraseness and informativeness are the two main features of keyphrase. Phraseness is a fairly dynamic idea which depicts the degree to which a given word sequence is considered to be a phrase. Informativeness denotes how well a phrase catches or outlines the important notions in a set of documents. A set of keyphrases related to a document gives high- level description of a document content that helps readers in searching for relevant information.

Keyphrase extraction in a news web page has been a challenging research topic in recent years because news changes very rapidly. Only a small number of news websites have author given keyphrases and manually allocating keyphrases for each web news document is very effortful. Thus it is absolutely necessary to automatically extract keyphrases. Automatic keyphrase extraction benefits users for the large document collection. Keyphrases of a document should be semantically related with the other words of the document. Therefore, in this paper, we proposed a Keyphrase Extraction approach, which uses lexical chain of semantically related words that are interconnected by semantic relations. The number of words and the number of semantic relations among the words can be different for each lexical chain. WordNet is used for the construction of lexical chain.

The organization of the paper is follows as. In section 2, previous studies on keyphrase extraction are discussed first. Section 3 describes the dataset used in the experiment. In section 4, we describe our proposed approach for the news web page keyphrase extraction. Experimental results and evaluation are discussed in section 5. Finally, some concluding remarks and future scope is discussed in section 6.

2. Related Work

In the previous works authors have suggested that document keyphrase can be useful in many areas as information retrieval and summarization. Chien [1] developed a keyphrase extraction system for Chinese and other Asian languages. Witten I. H. et al. [2] describe KEA algorithm, based on Naïve Bayes classifier automatically extracts keyphrases from text. This algorithm recognizes candidate keyphrases using lexical methods and computes feature values for each candidate by using machine learning algorithm and analyze which candidates are noble keyphrases. Martinez J. L. et al. [3] focus on AKE (Automatic Keyword Extraction), it is a keyword extraction system which is used to extract news articles keywords. KIP (Keyphrase identification program) [4] uses sample human keyphrsaes and then learns to identifiy additional news keyphrases. KIP mines noun phrases from documents and score will be allocated to each noun phrases. Depending on the weights the words that have higher score than the threshold will be selected as keyphrases. Wang J. et al. [5] proposed in their paper Neural Network based keyphrase extraction method. Lui Y. J. [11] presents a domain independent keyphrase extraction algorithm, which distinguish keyphrases from non – keyphrases by using statistical and computational linguistics techniques combination, a new attribute set and a new machine learning method; and shown that it perform well than other keyphrase extraction methods. Li z. F. et al. [12] proposed an approach based on lexical chain by using Reget's thesaurus and improve the KEA keyphrase extraction. Duwairi R. et al. [13] presents a framework for keyphrase extraction based on the KEA system. It relies on supervised learning particularly Naïve Bayes algorithm. Xu, S. et al. [14] introduce several novel word features by extracting inlink, outlink, category and infobox information from Wikipedia article set. Luo, Z. et al. [15] propose a method to integrate the comment posts for keyphrase extraction from web news documents. Boudin, F. [16] present and compare five centrality measures for graph based keyphrase extraction and used three datasets of different language and domain. Their results outperform the other centrality measure on short documents. Xie, F. et al. [17] proposes an approach which acquires semantic features within phrases from a single document. Their result demonstrates better performance than TFIDF and KEA. Gao, Y. et al. [18] propose a method to extract hot keyphrases from news report; their method consists a two-step process of keyphrase extraction based on TF*PDF. In their method each step uses position- weighted TF*PDF schema. Li, Z. et al. [19] propose a method based on the lexical chain to improve KEA keyphrase extraction, their

experiments result shows improvements compare with KEA and Nguyen and Kan's method. Hsu, H. M. et al. [20] propose subject- keyphrase concept to extract subject-keyphrases from a documents. They use definition-use chain for subject-keyphrase extraction algorithm. Wang, C. et al [21] propose a system for automatic online news topic keyphrase extraction, their system perform effectively with 70.61% precision and 67.94% recalls.

3. Description of The Dataset

The online news articles have been chosen from the 'The Hindu' news website. All these selected news is world news posted from 20 April 2016 to 30 April 2016. Our dataset contains 150 web news documents. The key purpose, we select 'The Hindu' news website for the experiment is that every news web page has author assigned keywords. We take the author assigned keywords as gold standard keyphrase. We choose some keyphrase manually for each document. Most of the keyphrases consists of one or more than one words.

Keyphrases having more than three words are less in number in our dataset. Average number of manually assigned keyphrases per document is 15. Here it is interesting to note that all author allotted keyphrase for a document may not occur in the title of the document. Total number of noun phrases in our dataset is 2250. The total number of author assigned keyphrases for all the documents in our dataset is 479.

4. Proposed Method

In the proposed method firstly in the document words are segmented, stemmed and stop words are removed. After that candidate phrases from the document are identified. Weight of each candidate phrase is computed by the features TF*IDF, phrase distance, and building lexical chain. According to the weight, a high scorer candidate phrases is selected as a keyphrases. The process of keyphrase extraction is shown in Fig1.

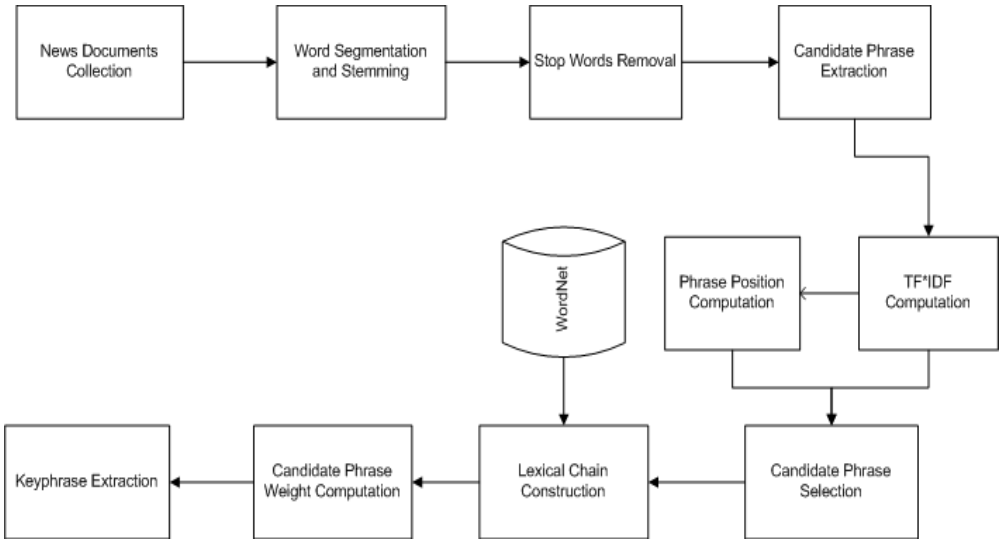


Fig.1. Keyphrase Extraction Process

The steps of the proposed method are as follows:

1. Words are segmented and stemmed and stop words are removed.
2. Identify the candidate phrase from each document.
3. Compute the TF*IDF and phrase distance of each candidate word

4. Select the top n candidate phrase according to the value of TF*IDF and phrase distance.
5. Build the lexical chains of each top n candidate phrase.
6. Compute the weight of each candidate phrase.
7. Select the top m candidate words as the keyphrase according to their weights. Select those candidate words as keyphrases which have higher weights.

4.1. Identification of Candidate Phrase

Keyphrases are extracted from candidate phrases. The noun phrases in the document are treated as the candidate keyphrase [6]. In order to recognize the noun phrases documents have been tagged by stanford Part-Of- Speech (POS) tagger [27]. We used Stanford POS tagger to extract the lexical information about the terms in a document. Fig. 2 shows the lexical tag assigned by the tagger for a document. According to this figure, JJ, DT, NN, NNS, VBZ, NNP, PRP$, VBN, IN, CD, etc are lexical tags assigned by the POS tagger.

Fig 3 shows the meaning of these tags. Candidate keyphrase extracted from Fig 2. Are: terrorists, central forensic science laboratory, DNA sample, government officials, National Investigation agency, investigation team, spokesperson, photographs, sensors.

Three|CD months|NNS after|IN Pathankot|NNP airbase|NN attacked|VBD terrorists|NNS belonging|VBG Pakistan-based|JJ Jaish-e-Mohammad|JJ (|NN JeM|NNP)|NNP forensic|JJ report|NN established|VBD six|CD terrorists|NNS present|JJ forensic|JJ examination|NN samples|NNS collected|VBN Billet|NNP last|JJ leg|NN operation|NN conducted|VBD found|VBN samples|NNS belonged|VBD "two|CD possible|JJ extract|NN any|DT "DNA|NN samples"|NN debate|NN number|NN terrorists|NNS present|NN airbase|NN came|VBD under|IN attack|NN alleged|VBD JeM|NNP terrorists|NNS intervening|VBG night|NN January|NNP 1|CD 2.|CD bodies|NNS terrorists|NNS killed|VBD within|IN first|JJ 24|CD hours|NNS operation|NN found|VBD airbase|NN bodies|NNS other|JJ found|NN Airmen's|NNS Billet|NNP where|WRB hiding|VBG blown|IN National|NNP Security|NNP Guard|NNP (|NNP NSG|NNP)|NNP different|JJ samples|NNS collected|VBN Billet|NNP Central|NNP Forensic|NNP Science|NNP Laboratory|NNP (|NNP CFSL|NNP)|NNP Chandigarh|NNP "The|NNP forensic|JJ report|NN says|VBZ samples|NNS tested|VBN positive|JJ human|NN remains|VBZ The|DT remains|NNS collected|VBD different|JJ rooms|NNS report|NN says|VBZ belong|JJ different|JJ humans|NNS remains|NNS badly|RB charred|VBN possible|JJ extract|NN DNA|NNP samples|NNS said|VBD senior|JJ government|NN official|NN National|NNP Investigation|NNP Agency|NNP (|NNP NIA|NNP)|NNP send|NN two|CD reminders|NNS CFSL|NNP before|IN reports|NNS sent|VBN February|NNP 8|CD The|DT Hindu|NNP giving|VBG details|NNS operations|NNS published|VBD interview|NN NSG|NNP Director-General|JJ R.C|JJ Tayal|NNP said|VBD certain|JJ six|CD terrorists|NNS airbase|NN sensors|NNS a|DT listening|NN device|NN wall|NN Airmen's|NNS Billet|NNP intercepted|VBD chatter|NN terrorists|NNS contacted|VBN Tuesday|NNP Mr.|NNP Tayal|NNP said|VBD "I|JJ seen|NN forensic|JJ report|NN always|RB said|VBD six|CD terrorists|NNS spokesperson|NN NIA|NNP said|VBD "|JJ want|NN comment|NN NIA|NNP preserved|VBD bodies|NNS four|CD terrorists|NNS shared|VBD photographs|NNS Pakistan|NNP through|IN letter|NN rogatory|NN Special|JJ Investigation|NN Team|NNP Pakistan|NNP expected|VBD visit|NN India|NNP conduct|NN joint|NN investigations|NNS

Fig.2. POS Tagged Document

CD: Cardinal number, DT: Determiner, NN: Noun (Singular or mass), VBZ: Verb (3rd person singular present), To: to, VB: verb (base form), VBN: Verb (past participle), RP: Particle, IN: Preposition, NNP: Proper Noun, NNS: Noun (Plural), CC: Coordinating Conjunction, VBG: Verb (gerund), WP: Wh- pronoun, JJR: Adjective, Comparative, NNPS: Proper Noun (plural), PRP$: Possessive pronoun, VBP: Verb (non-3rd person singular

Fig.3. Meanings of the Tags

4.2. TF*IDF of Candidate Phrase

After identifying candidate phrase, the collection of candidate phrases identified in the web news documents may be huge in number. From a vast collection a small number of phrases may be selected as the keyphrases. In this paper we select 15 keyphrases from a single document. TF*IDF of each candidate phrase is used to rank the phrases. TF*IDF measure the phrase frequency in a document compared to its rarity in general use.

We compute the TF*IDF of each word by the given eq. (1)

$$TF * IDF = \frac{t_f}{t_n} * \log(\frac{N}{n_i}) \tag{1}$$

Where tf is the frequency of term t in a document, tn is the total number of terms in a documents, N is the total number of documents and ni is the number of documents in the dataset that contains term t.

4.3. Phrase Distance

The distance attribute is the position where a phrase first appears in the document. The candidate keyphrases that appears early in a document should be given higher score. Like previous approach [7], Distance of a phrase from the start of a document is measured as the number of words that precede its first appears divided by the number of words in the documents. The distance of a phrase in the document is calculated as in eq. (2)

$$PhraseDis\tan ce = \frac{n_j}{n} \tag{2}$$

Where nj is the number of words that predate its first appearance, and number of words in the document are denoted by n.

4.4. Construction of Lexical Chain

Firstly Morris and Hirst [8] give the concept of lexical chain. According to them lexical cohesion is an arrangement of related words that give the continuity of lexical meaning. Lexical cohesion occurs as a result of semantic relation between words. One of the main advantages of lexical cohesion is that it is an easily recognizable relation that enables the computation of lexical chain. Lexical chains visualize the semantically related words or phrases in the text. These words or phrases are called the lexical items and each item gives a specific meaning to a lexical chain. In this paper we use WordNet for creating lexical chains. With the help of path between concepts, lexical chain can be found. In general two concepts can have many possible lexical chains. For creating lexical chains we ignore numbers, units, currencies, times/periods, names, places and referring items [10]. Suo, Hong-guang et al. [22] use HowNet to determine the relationship between words and build vocabulary chain. For the construction of lexical chain we used synonym, hypernym/hyponym, coordinate term and meronym, Silber, H. Gregory [23] and Ercan, Gonenc [9] also used the same relations except coordinate term. In order to rank lexical chains, high scoring chains must be picked as the important concept from the original document. We use Barzilay and Elhadad [24] idea of strong chain.

Fig 4 shows the different set of lexical chains chooses from the tagged document in Fig 2.

Lexical chains usually depend on semantic relations that can be acquired from WordNet. Hypernym/ Hyponym, Synonym/ Repetition, Meronym/ Holonym, Antonym, and Sibling relations are used to build lexical chains.

Fig 5 shows the lexical graph of LC1 in detail.

Weights of every relation between word senses are given allegedly [9]. Table 1, shows the allocated weights for the relation. Subsequent to scoring each lexical chain of the word, we select the chain with a maximum score as the lexical chain.

Table 1. Weight of Lexical Chain Relation

Relation	Explanation of Relation	Weight
Synonym/ reiteration	Same meaning	10
Coordinate Term	Sibling	8
Hypernym/hyponym	General/specific	7
Meronym	Is a part of	4

According to these assigned weights, the score of lexical chain LC1 is equal to 43 (=5*7 + 8) since there are five Hypernym/Hyponym relations and one Coordinate term.

LC_1= {terrorists, terrorists attack, Pakistan-based JeM, JeM}
LC_2= {Central forensic science laboratory, forensic examination sample, forensic report, DNA sample}
LC_3= {Government officials, National Investigation Agency, Investigation Team}
LC_4= {Spokesperson, Photograph, Sensors}

Fig.4. Set of Lexical Chain

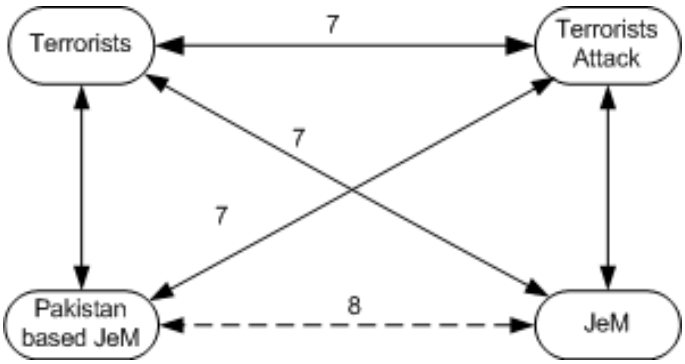


Fig.5. Lexical graph

4.5. Weight of Candidate Phrase

Weight of a candidate phrase can be obtained by the combination of the features: TFIDF, phrase distance, and lexical chain shown in eq. (3).

$$\text{weight} = a * TF * IDF + b * \text{phrase distance} + c * \text{Lexical chain} \quad (3)$$

Where TF*IDF is the value of the candidate phrase, phrase distance is the distance into the document of candidate phrase first appearance, and lexical chain is the length of the chain that contains candidate phrase and a, b, c are the parameters that can be adjusted. The value of these parameters in our experiment has been set to 1.

Suppose we have to find the weight of a candidate phrase "Terrorist", which we selected from our dataset. The weight computation as discussed in eq. (3) comprise of three components, the value of each component has been obtained as follows.

Firstly we calculate the TF*IDF of the phrase "Terrorist". The position of "Terrorist" phrase in the document of our dataset is 9 and the number of words in the whole document is 367, then the value of TF*IDF is calculated as in eq. (1)

$$TF * IDF = \frac{t_f}{t_n} * \log(\frac{N}{n_i})$$

TF*IDF= 0.0174832

Secondly we find the value of Phrase distance of the "Terrorist" phrase as calculated in eq. (2)

$$PhraseDis\tan ce = \frac{n_j}{n}$$

Where the value of nj is 6 in the document and n is 367. Therefore the value of phrase distance is

Phrase distance = 0.0163

In the next step, we construct and then calculate the value of lexical chain. We choose lexical chain LC1 from fig. 4, because the value of LC1 is higher than other lexical chains. The value of lexical chain LC1 is 43, calculated in previous section 4.4. Finally the weight of the candidate phrase is calculated as:

W= 1*0.0174832+ 1* 0.245 + 1* 43
W= 43.0174832 ~ 43.01

Like the "terrorist" phrase, all the candidate phrases of the dataset are calculated. We select the top fifteen higher weight scorer candidate phrases as keyphrases of a document.

5. Experiment Result and Evaluation

Experiments were carried out to evaluate the overall performance of our approach. For evaluating the automatically generated keyphrases, we first take the two standard information retrieval metrics precision and recall. The precision; measures the proportion of number of extracted key phrases that are also author tagged key phrases to the total number of extracted keyphrases. The second one 'recall' measures the proportion of the extracted key phrases that are also author tagged key phrases to the number of author tagged keyphrases. These metrics show how well generated phrases match a set of relevant phrases.

$$\Pr ecision = \frac{N_{e \cap t}}{N_e} \tag{4}$$

$$\mathrm{Re}call = \frac{N_{e \cap t}}{N_t} \tag{5}$$

Where Ne is the number of keyphrases extracted, Nt the number of keyphrases tagged by author. Ne∩t is the number of extracted keyphrases that are also keyphrases tagged by author.

Table 2 shows the keyphrases assigned by the author of the news article which is the document number 2 in our dataset.

Table 2. Author Assigned Keyphrases for News Article Number 2 in the Dataset

Document No.	Author Key
2	Pathankot attack
2	forensic attack
2	Terrorism
2	Special Investigation Team
2	Joint investigation

From the document 2, our proposed approach extracted the top 5 keyphrases as shown in Table3.

Table 3. Top 5 Keyphrases Extracted By our Proposed Approach

Document No.	Author Key
2	Pathankot attack
2	forensic science laboratory
2	terrorism
2	Special Investigation Team
2	National Investigation agency

Table 2 and Table 3 show that out of 5 keyphrases extracted by our approach, 3 keyphrases matched with the author assigned keyphrases.

In order to compare our approach with state- of- the-art keyphrase extraction systems we have selected KEA [2] and KESR [25]. Most existing systems identify candidate phrases by the method applied in KEA and KESR.

KEA is comparatively simple and useful in automatic keyphrase extraction. The KEA identifies candidate keyphrase using lexical methods and calculates the feature value of each candidate phrase, and then predicts the good keyphrase from candidate by using machine learning algorithm. The basic model of KEA involves two stages. Firstly build a model for recognizing keyphrases by using training documents where the author keyphrases are known. Secondly, use the model create on first stage, choose the keyphrases from a new document. The overall performance of KEA show that on average KEA can match between one and two of the five keyphrases chosen by the average author in the collection.

NFAS system considers all phrases except stop words in the web news pages. In this system Key-phrase Extraction based on Semantic Relations (KESR) algorithm is used for keyphrase extraction. The goal of KESR is to extract those words that have a low frequency but provide a major impact to the text subject. The basic model of KESR algorithm involves two attributes: TFIDF, and word similarity and lexical chain. Word similarity is computed through HowNet. Extracted keyphrases compared with the phrases in the news title and phrases in the core hints provided by the author. By comparing their results with TF*IDF and KELC (Key-phrase extraction based on lexical chains) [22], KESR outperforms the other two in both the cases, when the title kept and when the title removed and core hints kept.

In this paper, we compare the overall performance of our keyphrase extraction method with the existing keyphrase extraction methods. In the experiment, the number of keyphrases to be extracted was set to 5, 10, and 15 respectively. Table 4, shows that the approach presented in this paper seems to be better than other approaches in terms of precision and recall.

Table 4. Precision and Recall Comparison of Three Approaces

Number of Keyphrases	Average Precision			Average Recall		
	Our Approach	KESR	KEA	Our Approach	KESR	KEA
5	0.34	0.32	0.28	0.25	0.24	0.29
10	0.22	0.20	0.19	0.46	0.36	0.40
15	0.17	0.18	0.15	0.51	0.41	0.48

Fig 6 shows the comparison of the individual performance of three different approaches. Precision is the proportion of the keyphrases extracted that are correct. The experiments indicates that the precision of our approach when extracting 5, 10, and 15 keyphrases is 0.34, 0.22 and 0.17 respectively is greater than KEA and KESR for the same 5, 10 and 15 keyphrases 0.28, 0.19 and 0.15; and 0.32, 0.20, and 0.18 respectively.

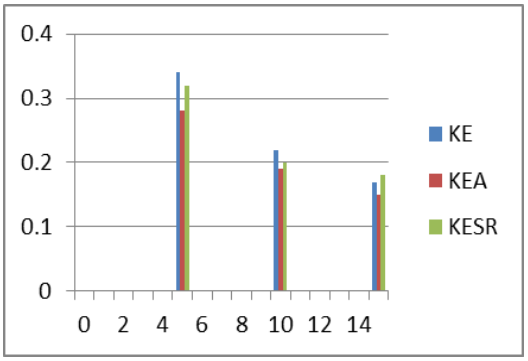


Fig.6. Precision Comparison of Three Algorithms.

Fig 7 shows the recall comparison of three different approaches. Recall is the fraction of relevant instances that are retrieved. Recall of our approach when extracting 5, 10, and 15 keyphrases is 0.25, 0.46 and 0.51 respectively is greater than KEA and KESR for the same 5, 10 and 15 keyphrases 0.29, 0.40 and 0.48; and 0.24, 0.36, and 0.41 respectively.

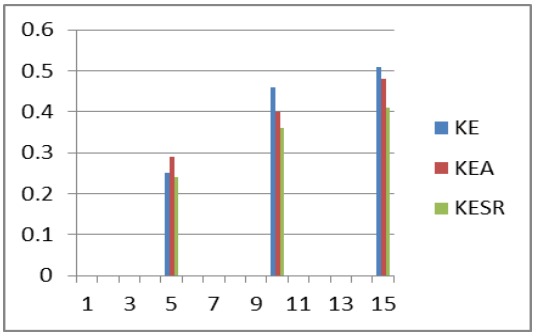


Fig.7. Recall Comparison of Three Algorithms

6. Conclusion and Future Work

In this paper, we presented an effective technique which can extract keyphrases from news web page. In this work, we take noun phrases of the documents as a candidate Phrase, and used POS tagger for this task. While ranking candidate keyphrases, weights of each candidate keyphrase is measured and choose the highest score keyphrases. To determine the weight of the keyphrase we use the TFIDF, phrase distance in the document and lexical chain. The approach is evaluated by the evaluation parameters precision and recall. Experimental results show that this approach is competitive with other known approaches.

In the future, we might want to use more datasets to assess our system. For keyphrase extraction algorithm there is no standard datasets are available. In this paper we compare the extracted keyphrase with the author assigned keyphrases. But there are many other issues in this method. First, author assigned keyphrases are not generally show up in the document to which they belong. So, if a keyphrase is not contained in a given web page, it is never extracted as a keyphrase of the given web page by the automatic keyphrase extraction algorithm. Second authors provide only limited keyphrases which are less than extracted automatic. Therefore we will accomplice more research on searching for a more logical and target approach to assess the automatic extraction results.

References

[1] Chien LF. PAT-tree-based adaptive keyphrase extraction for intelligent Chinese information retrieval. Inf. Process. Manage.1999 Jul 1; 35(4):501-21.

[2] Witten IH, Paynter GW, Frank E, Gutwin C, Nevill-Manning CG. KEA: Practical automatic keyphrase extraction. InProceedings of the fourth ACM conference on Digital libraries 1999 Aug 1; 254-255.

[3] Mart ínez-Fern ández JL, Garc á-Serrano A, Mart ínez P, Villena J. Automatic keyword extraction for news finder. InInternational Workshop on Adaptive Multimedia Retrieval 2003 Sep 15; 99-119.

[4] Wu YF, Li Q, Bot RS, Chen X. KIP: a keyphrase identification program with learning functions. In Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on 2004 Apr 5; 2: 450-454.

[5] Wang J, Peng H, Hu JS. Automatic keyphrases extraction from document using neural network. Advances in Machine Learning and Cybernetics. 2006; 633-41.

[6] Wu YF, Li Q. Document keyphrases as subject metadata: incorporating document key concepts in search results. Information Retrieval. 2008 Jun 1; 11(3):229-49.

[7] Frank E, Paynter GW, Witten IH, Gutwin C, Nevill-Manning CG. Domain-specific keyphrase extraction. In16th International Joint Conference on Artificial Intelligence (IJCAI 99) 1999; 2: 668-673).

[8] Morris J, Hirst G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational linguistics. 1991 Mar 1; 17(1):21-48.

[9] Ercan G, Cicekli I. Using lexical chains for keyword extraction. Information Processing & Management. 2007 Nov 30; 43(6):1705-14.

[10] Steffen R. Lexical chain Annotation Guidelines. 2012.

[11] Lui YJ, Brent R, Calinescu A. Extracting significant phrases from text. In Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on 2007 May 21; 1: 361-366.

[12] Li ZF, Zhao XH, Yi J, He B. Improvement of KEA Based on Lexical Chain. In Advanced Materials Research 2013; 756: 2999-3004.

[13] Duwairi R, Hedaya M. Automatic keyphrase extraction for Arabic news documents based on KEA system. Journal of Intelligent & Fuzzy Systems. 2016 Jan 1; 30(4):2101-10.

[14] Xu S, Yang S, Lau FC. Keyword Extraction and Headline Generation Using Novel Word Features. In AAAI 2010 Jul 5; 1461-1466.

[15] Luo Z, Tang J, Wang T. Improving keyphrase extraction from web news by exploiting comments information. InAsia-Pacific Web Conference 2013; 140-150.
[16] Boudin F. A comparison of centrality measures for graph-based keyphrase extraction. In International Joint Conference on Natural Language Processing (IJCNLP) 2013; 834-838.
[17] Xie F, Wu X, Hu X. Keyphrase extraction based on semantic relatedness. In Cognitive Informatics (ICCI), 2010 9th IEEE International Conference on 2010; 308-312
[18] Gao Y, Liu J, Ma P. The hot keyphrase extraction based on tf* pdf. InTrust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on 2011 Nov 16; 1524-1528
[19] Li Z, He B. Adding Lexical Chain to Keyphrase Extraction. InWeb Information System and Application Conference (WISA), 2014 11th 2014; 254-257.
[20] Hsu HM, Chang RI, Chang YJ, Lin SY, Wang YJ, Ho JM. Subject-Keyphrase Extraction Based on Definition-Use Chain. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015 IEEE/WIC/ACM International Conference on 2015 Dec 6; 3: 199-202.
[21] Wang C, Zhang M, Ru L, Ma S. An automatic online news topic keyphrase extraction system. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 2008 Dec 9; 214-219.
[22] Suo H, Liu Y, Cao S. A keyword selection method based on lexical chains. Journal of Chinese Information Processing. 2006; 20(6):25-30.
[23] Silber HG, McCoy KF. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. Computational Linguistics. 2002 Dec; 28(4):487-96.
[24] Barzilay R. *Lexical chains for summarization* (Doctoral dissertation, Ben-Gurion University of the Negev). 1997
[25] Wu X, Wu GQ, Xie F, Zhu Z, Hu XG. News filtering and summarization on the web. IEEE Intelligent Systems. 2010 Sep; 25(5):68-76.
[26] http://nlp.stanford.edu/software/tagger.shtml

Authors' Profiles

Chandrakala Arya is Research Scholar at Department of Computer Science in B.B. Ambedkar University, Lucknow, India. She has received her M.C.A. Degree in the year 2011 from Uttarakhand Technical University. Her research interest includes Information Extraction and Text Summarization. She has published some of the research papers in international conferences.

Prof. Sanjay K. Dwivedi is Professor and Head at Department of Computer Science in B.B. Ambedkar University, Lucknow, India. He has received his Ph.D. Degree from Banasthali Vidyapeeth in area of Web Mining in the year 2006. His research interest includes Web content Mining, Semantic Web, Search Engine performance evaluation, Machine translation, Information Retrieval etc. He has published many of the valuable research papers in various national and international Journals of repute.

How to cite this paper: Chandrakala Arya, Sanjay k. Dwivedi,"Keyphrase Extraction of News Web Pages", International Journal of Education and Management Engineering(IJEME), Vol.8, No.1, pp.48-58, 2018.DOI: 10.5815/ijeme.2018.01.06

Content extraction from news web pages using tag tree

Chandrakala Arya* and Sanjay K. Dwivedi

Department of Computer Science,
Babasaheb Bhimrao Ambedkar University,
Lucknow, India
Email: arya.chandrakala@gmail.com
Email: skd200@yahoo.com
*Corresponding author

Abstract: As the web endures to develop, there is an enormous amount of information which is typically designed for its users, which makes it difficult to extract relevant data from numerous sources. In this paper, we propose an approach for extracting the main content from news web pages. Our approach is based on the concept of tokenisation of HTML page, these tokens construct the tag tree; web pages from different websites are parsed into Tag tree and generated a template from each web pages and discover matching patterns and multiple sequence alignment. It finds and removed shared token sequences from the web pages until the relevant information is extracted from them. We perform experiments on 500 web pages from ten different news websites. Experimental results show that our approach efficiently extracts the relevant information.

Keywords: content extraction; tag tree; news web page; information extraction; pattern matching.

Reference to this paper should be made as follows: Arya, C. and Dwivedi, S.K. (2018) 'Content extraction from news web pages using tag tree', *Int. J. Autonomic Computing*, Vol. 3, No. 1, pp.34–51.

Biographical notes: Chandrakala Arya is a Research Scholar at the Department of Computer Science in the BB Ambedkar University, Lucknow, India. She received her MCA degree in 2011 from the Uttarakhand Technical University. Her research interest includes information extraction and text summarisation. She has published some of the research papers in international conferences.

Sanjay K. Dwivedi is a Professor and Head at the Department of Computer Science in the BB Ambedkar University, Lucknow, India. He received his PhD degree from the Banasthali Vidyapeeth in the area of web mining in 2006. His research interest includes web content mining, semantic web, search engine performance evaluation, machine translation, information retrieval, etc. He has published many of the valuable research papers in various national and international journals of repute.

1 Introduction

There is a vast amount of information available on the web, but most of the information is not in a form that can be easily used by the user. Efficient access to the relevant information within the huge amount of information needs major efforts. The process of content extraction is defined as the extraction of relevant content from massive data, such as text, database, semi-structured and multimedia documents. Efficiently extracting high quality content from news web pages is a challenging yet important problem in the field of information retrieval and news summarisation. News articles are unstructured documents, whose relevant information is pieces of free text. To extract the relevant news from the whole web page, our approach identifies and searches common characteristics that are usually present in the news web pages. Like most news websites have the following structure:

1 a home page that presents the important headlines from all fields

2 several section pages divided in different areas of interest like business, sports, national, international, technology, etc., that provide the related headlines

3 pages that actually present the news, containing the title, author, date and body of the news.

Our approach is based on the basic assumption that the news web pages content is divided into tokens where tokens represent the HTML tags like <head><title><script>, etc.

A major problem in news content extraction is mining useful information from web because news web pages not only contains the actual news content but also some noisy content like advertisement, comments and branding banners, etc. Therefore in the news web page, the actual news content is just half, and noisy content occupies nearly half of the page. We faced the problem of identify core content even during processes that includes human assessment. In our work we extracted the core content from a large number of news web pages and these news web pages comes from ten different news websites. We mainly deal with news pages written in English.

Identification of actual news content from news web pages is relatively easy task for the human being, who can identify just by visual inspection; however it is hard problem for machines. For content extraction there have been many existed approaches.

Our approach not only extract the relevant text passage from the given news website but also the fetching of the entire website content, and the extraction of the relevant content.

Content extraction that depend on extraction rules do not usually adapt well to changes to the web. When the set of extraction rules is handcrafted or learnt, the web keeps growing and it is uncommon that changes may invalidate the existing extraction rules. Therefore some authors work on semi-automatic extraction rules. Our work does not rely on extraction rules like previous approaches. It requires input web pages and translated into tag tree. It works on two or more web documents and compares them to obtained shared patterns that are likely to provide relevant information. The idea of

identifying shared pattern relies on tree matching and determines which are equivalent to one another and apply filtering algorithm to filter out irrelevant content. We have conducted experiments with 500 news web pages from ten different news websites and our results confirms that our approach can achieve precision as high as 96% and recall as high as 97%.

The rest of the paper is organised as follows: Section 2 presents the related work in the field of information extraction. Section 3 describes our proposed approach. Dataset used in the experiment is discussed in Section 4. Experiment and results are discussed in Section 5. Finally we conclude our work in Section 6.

2 Related work

Plenty of work has been reported in literature for content extraction from news web pages including standard and independently techniques. Now a day many researchers paying attention in the field of content extraction from news web pages. Among the variety of work reported in literature, we reviewed the various other techniques that are attempting to solve the similar problem.

In 2007, Banko et al. introduces a fully implemented, highly scalable OIE system TEXTRUNNER, which has the ability to extracts large amount of high quality data from a nine million web page corpus. In this system tuples are assigned a probability and indexed to support efficient extraction and exploration via user queries. They compare TEXTRUNNER with KNOWITALL, and achieve an error reduction of 33%.

Knoblock et al. (2003) developed a set of tools for extracting data from web sites and transforming it into a structured data format such as XML. Their approach automatically detecting the breakage of wrapper and repairing them capitalises on the regular structure of the extracted fields themselves. Their technology learns highly accurate extraction rules and wrapper is verified by the correct extraction of data.

Lin and Ho (2002) propose InfoDiscoverer system for the identification of informative contents from a web page. Their system partition the web page into several content blocks according to HTML tag <Table>. Their proposed method dynamically select the entropy-threshold that partition the block into either informative or redundant. Their experimental results show the value of precision and recall greater than 0.956.

In 2009, Louvan propose an approach for the extraction of main content from web documents by using the combination of machine learning and heuristics approaches. Their results show that the combination of classification task and LBS namely CS+LBS gives best performance for blogs and news datasets.

Prasad and Paepcke in 2008 developed a heuristic technique for the extraction of main contents of a news web page. They construct DOM tree of the web page and scored the nodes according to amount of text and number of links it contains. There method is site-independent and does not used any language based features. The performance of their algorithm achieved 97% precision and 98% recall which is slightly below the baseline system but requires significantly less computational speed; the processing speed of algorithm is less than 15 ms per page.

Gibson et al. (2007) used machine learning methods to identify the content of a news web page. They identify correctly the portion of the content of web pages from

80%–97% of the time. Their experimental result shows that the document level accuracy of their model is 80% and also shows the high value of precision and recall for content block identification.

Reis et al. in 2004 present a domain oriented approach to web data extraction. Their approach of data extraction from web pages is based on the structure analysis of the target web pages. The structure of the web page can be described by a DOM tree, since they introduced a new algorithm RTDM for calculating the edit distance between two given trees and solve the problem of structure-based page classification, extractor generation and data labelling. Their experiment results show that RTDM correctly extracting 87.71% of news from a dataset of 4,088 pages.

Gupta et al. (2005) developed a framework that use DOM trees and a W3C specified interface that allows programs to dynamically access the structure of a document. Their approach is implemented in a publicly available Web proxy for the content extraction from HTML web pages.

Ji et al. in 2010, proposed web information extraction method that is based on Tag tree template and efficiently extract meaningful information including records and data schema. They describe a web page by HTML Tag tree and both HTML tag and text are treated as tree nodes. Their result shows the effective extraction of meaningful information.

Sleiman and Corchuelo in 2013 propose an unsupervised information extraction approach TEX, works on the idea that two or more web documents generated by the same server side template and removes shared token sequences among web documents until finding the relevant information that should be extracted from them. TEX working on malformed web documents and reduces extraction time by not converting HTML code into XHTML and DOM trees. There technique achieves a very high precision and recall value of approximately 100%.

Kaddu and Kulkarni in 2016 present a hybrid approach for the extraction of web page main contents. Their approach is based on the combination of automatic extraction and manual hand crafted rule techniques. They generate rules by machine learning method, by using that rules relevant content from web pages are extracted. Their work generates effective rules and achieves automaticity and efficiency. Their results show that retrieval accuracy depends on the size of the dataset. For a lower number of files it shows 28%–30% accuracy and for higher number of files it shows 90% retrieval accuracy.

Pettersson et al. in 2016 presents HistSearch tool for automatic information extraction from historic text. They present the outcome of collaboration between the field of computational linguistics and history, which resulting a graphical user interface for information extraction from historical text. They describe the workflow of the system, based on the spelling normalisation using advanced taggers and parsers available for the standard modern language. A prototypical graphical user interface used by the historians and a manual evaluation of the tool performed by the actual users. Their results show that spelling normalisation is successful for the task of tagging and lemmatisation.

Ziegler et al. in 2007 (2007) present an approach that works in a fully-automated fashion, classifying text blocks in html pages using distilling linguistic and structural features, having a particle swarm optimiser (PSO) learn feature thresholds for optimal classification performance. Their approach shows good results for hundreds of news pages from popular media in five languages and exhibits an accuracy that comes close to human judgment.

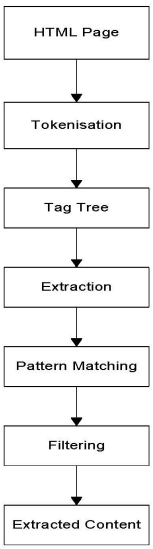
Muslea et al. in 1999 presents an approach to wrapper induction which is based on the idea of hierarchical information extraction. They introduce an inductive algorithm, STALKER that generates high accuracy extraction rules based on user-labelled training examples. Their experimental results show that STALKER does significantly better than other approaches.

Guo et al. (2010) propose an approach called ECON for content extraction from news web pages. The approach finds a snippet-node which is used to wrap the news content and then backtracks from the snippet-node until a summary node is found, and then the summary node wrapped the entire news content. In this approach backtracking removes the noise. Their experimental results show that ECON can achieve high accuracy of 90% for scalable extraction.

3 Proposed algorithm

Our algorithm is divided into four components. First one is tag tree that is used to measure the similarity between the templates of web pages. The second component is an extractor that returns a list of ContentSet that contains as much potential information as possible. The third component is pattern matching that finds out repetitive pattern. The fourth component is a filter that filters out undesired patterns and return candidate pattern. Our proposal is works on a collection of web documents, which we denote as ContentSet (CS). A content set is a set of contents that are sequences of HTML tags. Our implementation and our experiments were based on these HTML tags. We create a Tag tree to describe a web page, where tree nodes are defined by HTML tag and text. HTML tags are the basic components for document presentation and tags themselves convey certain structure information.

Figure 1 Process of content extraction



Referring to Figure 1, a flowchart of the content extraction process is shown. When a user submits a html page, the html page is tokenised into simple tokens represents either script blocks, style blocks or html tags. The Tag tree is constructed by these tokens. The pattern matching then uses the Tag tree to discover repetitive patterns. The repetitive patterns are forwarded to filtering, which filter out undesired patterns and finally the extracted content is found.

We present the algorithm that works on a collection of web documents.

Algorithm 1 Content extraction

```
1   Tt = Tag tree (CS, html tags)
2       extract ( CS: ContentSet; End, Start) : List <ContentSet>
3           l = extract (CS, Tt, End, Start)
4               m = PatternMatching (l)
5           Result = filter (m)
6   Return result
```

The algorithm works in four steps: at line 1, we invoke Tag tree. At line 2, we invoke algorithm extract, which makes an attempt to extract the information that varies from document to document. At line 4, we invoke the pattern matching algorithm. This algorithm searches for the shared patterns of size end, end-1, … … start. If start > 1 or

End is less than the size of the input document, then the search has a preference that may lead to situation in which pattern matching algorithm return information that actually belong to the template, therefore at line 5, we invoke filtering algorithm.

3.1 Tag tree

HTML in a web page is parsed as a Tag tree in our work. Tag tree could be thought as the base structure to implement DOM. Attributes in Tag tree are regarded as nodes. The child nodes and attributes of a tag tree can also be well sorted and indexed.

Following rules are used for the construction of Tag tree.

1 there are mainly three type of node in the Tag tree, which is summary node, text node and Tag node

2 entire content of news with its subtrees are wrapped in a pair of node such as <HTML></HTML> is summary node

3 tags and text between a pair of tags, such as <body></body> are all children nodes of the Tag tree

4 all the content between a pair of <script> tags is a text node as the only child of <script> node

5 all attributes of a node which are parsed in order, will be inserted into the attribute map

6 a tag ended with '/>' is a node with self-closing flag is true such as <frame.src= "sun.htm"> of a XHTML.

Figure 2 Source page (see online version for colours)

```
<!DOCTYPE html>
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en">
<head>
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=ecge,chrome=1">
<title>
Four killed as militants ambush Army convoy in Shopian - The Hindu
</title>
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<meta name="title" content="Four killed as militants ambush Army convoy in Shopian"/>
<link rel="shortcut icon" type='image/x-icon"
href="http://www.thehindu.com/favicon.ico"/>
<link rel="icon" type="image/ico" href="http://www.thehindu.com/favicon.ico"/>
<link rel="apple-touch-icon" href="http://www.thehindu.com/apple-touch-icon.png"/>
<link rel="apple-touch-icon-precomposed" href="http://www.thehindu.com/apple-touch-icon-
precomposed.png"/>
<meta name="msapplication-starturl" content="http://www.thehindu.com/"/>
```

Figure 3 Tag tree

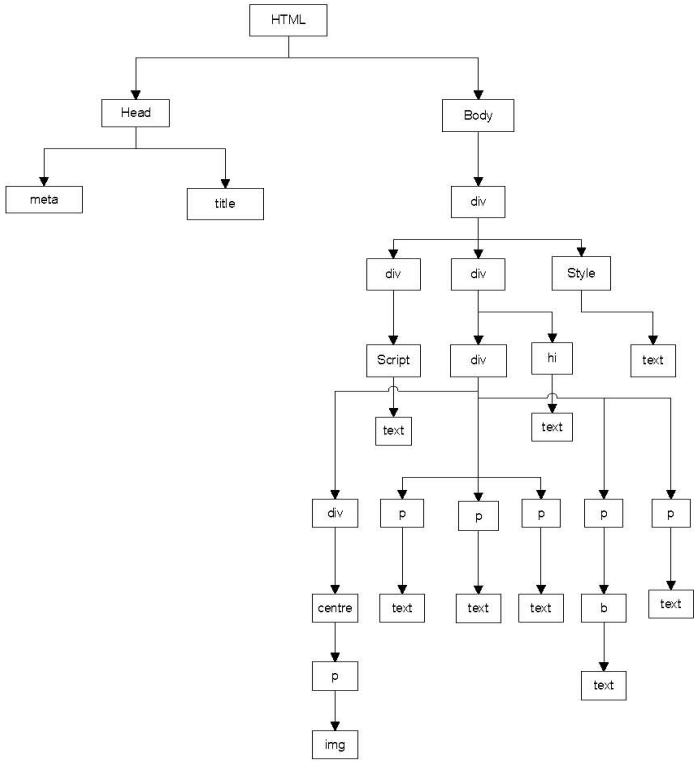


Figure 2 shows the news web page source page. We create Tag tree with the help of this source page.

Figure 2 shows the structure of a web page. By manual inspection we found that all news web pages show the similar structure and contains following tags <html> <head><meta><title></title><link ref><script></script> <p></head></html>. With the help of these tags we construct the Tag tree. Figure 3 shows the tag tree of a news web page.

3.2 Extract algorithm

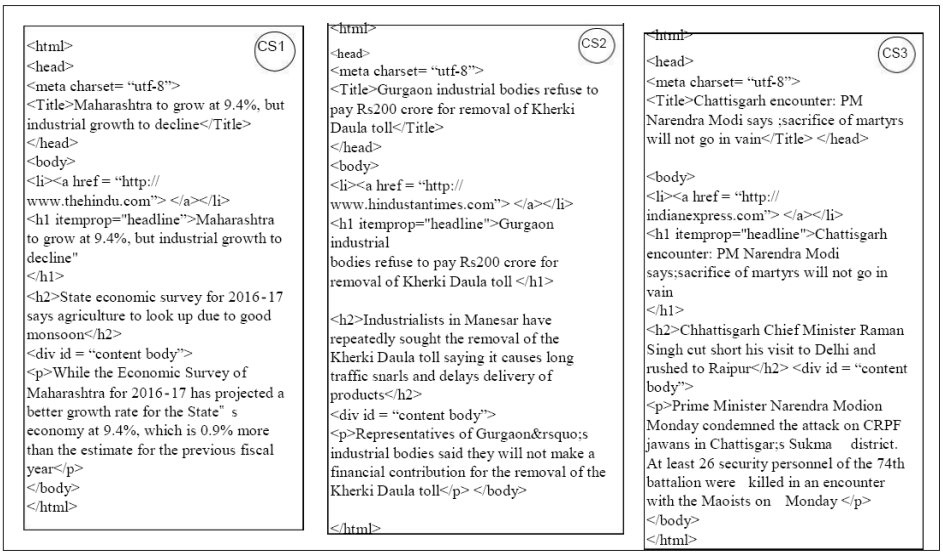
Extract algorithm searches for shared patterns of size start down to end in a Contentset. We find the shared patterns using Tag tree. Algorithm 2 presents the algorithm extract that lies on the contents, which makes an attempt to extract the information that varies from document to document. It works on a collection of web documents which we denote as ContentSet. Intuitively, a content set is a set of contents which are sequences of tokens. Content is not bound with a particular tokenisation schema. Our implementation and our experiment were carried out using a simple tokenisation schema according to which tokens represents Tag set and we construct a Tag tree as shown in Figure 3. The algorithm extract can search the tag tree to find all occurrences of the extraction pattern. In Figure 3 the longest shared pattern we used is of size 7 tokens <html><head><title>News title</title></head></body></html>. In this algorithm, the main loop at lines 3–15 iterates over all possible sizes from start down to end. The inner loop at lines 5–13 searches for a shared occurrence of that size.

Algorithm 2 Extract

```
1   Extract (CS : ContentSet ; Start, End, TT: Tag Tree) : List <ContentSet>
2           Result= <CS>
3       for Size = Start down to End do
4               Buffer = <>
5               while result ≠<> do
6                   CS = dequeue (result)
7               if TT= SharedOccurences(CS, size) then
8                   enqueue (buffer, CS)
9               else
10                  enqueue (result, TT)
11              end
12          end
13      Result= buffer
14  end
15  Return result
```

This algorithm of extract news is to extract entire content from web news page. The input of it is a web news page. The extracted document is shown in Figure 4.

Figure 4 Extract result



3.3 Pattern matching

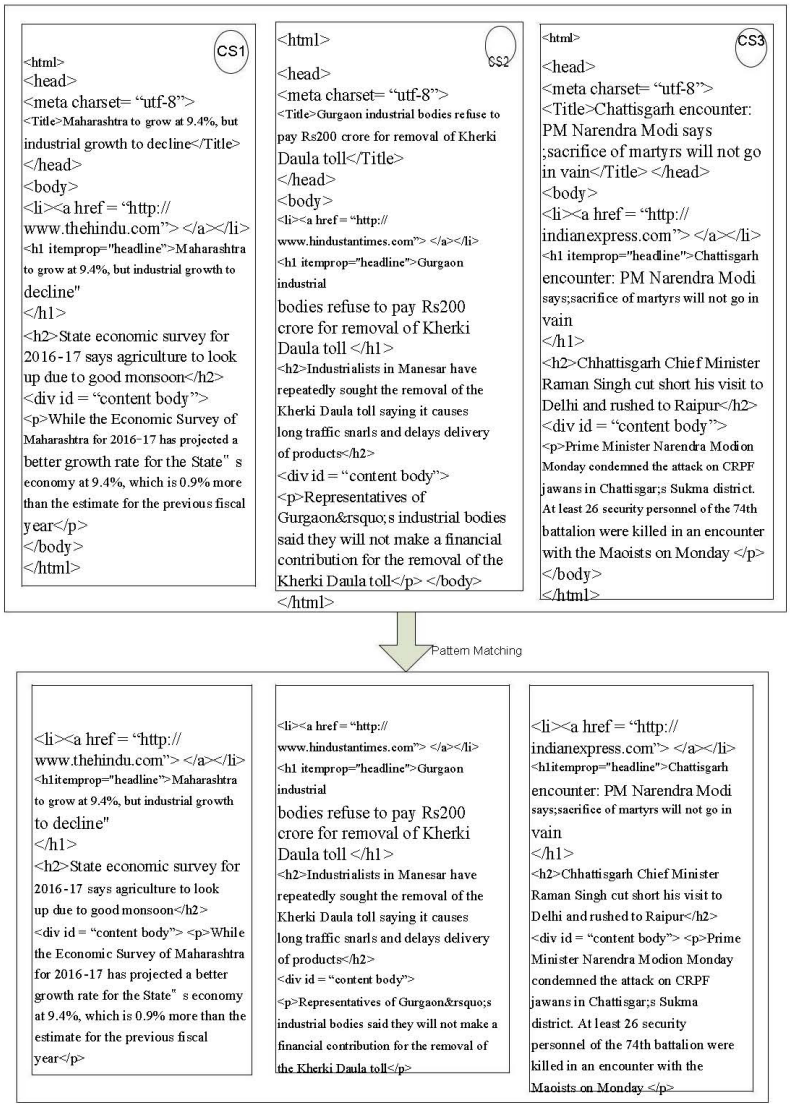
After the extraction, user may select target pattern that contain desired information. Algorithm 3 presents the pattern matching algorithm. It works on a ContentSet CS, a content base of size s, which is supposed to be the shortest non-empty content in CS. The goal of this algorithm is to find and match patterns inside the base that occurs in every content in CS. In this algorithm, at lines 3–11, i iterates from 0 until size s. as long as no matching pattern is found. The actual search is performed in the inner loop at line 6-10; in this loop the algorithm iterates over every content in the input ContentSet and finds all the matching patterns that start at position i and has size s. The algorithm returns a list of matching patterns in the ContentSet.

Algorithm 3 Pattern matching

```
1   PatternMatching (CS: ContentSet ; base: Content; TT: Tag Tree) : Map<ContentSet>
2       Found= false
3       For i = 0 until size (base) – s while not found do
4           Result = {}
5           found= true
6           foreach Content in CS while found do
7               If TT = findPatternMatches ( Content, base, s) then
8                   found = size (matches) > 0
9               else
10                  found = size (no matches)
11                  result= result ∪ {Content ↦ PatternMatches}
12              end
```

13 end
14 return result

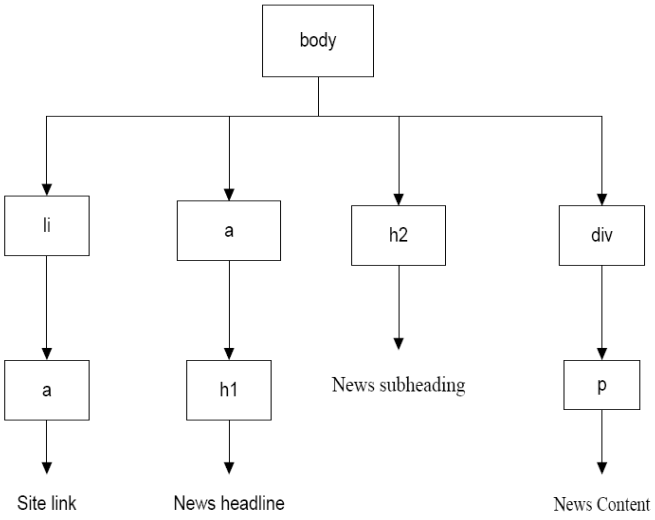
Figure 5 Searching for pattern matching using <li> as a base (see online version for colours)



The result of pattern matching is shown in Figure 5. All the leaves in a Tag tree share a common prefix, all the three news web pages shows the common Tag tree. Leaves represent the repeated sequence of input. In figure 5 to search for a pattern of size 4; we supposed that base is the shortest content in CS1, CS2, CS3. The algorithm first searches for site link
<ahref= “site link”>
in every content in CS1, CS2 and CS3 and found it. Then it searches for <h1= “headline”></h1>, <h2>sub headline</h2> and

 News content which is found in every ContentSet CS1, CS2, CS3. As a conclusion <a>, <h1>, <h2>, are the matching pattern in the content set.

Figure 6 Tag tree of pattern matching



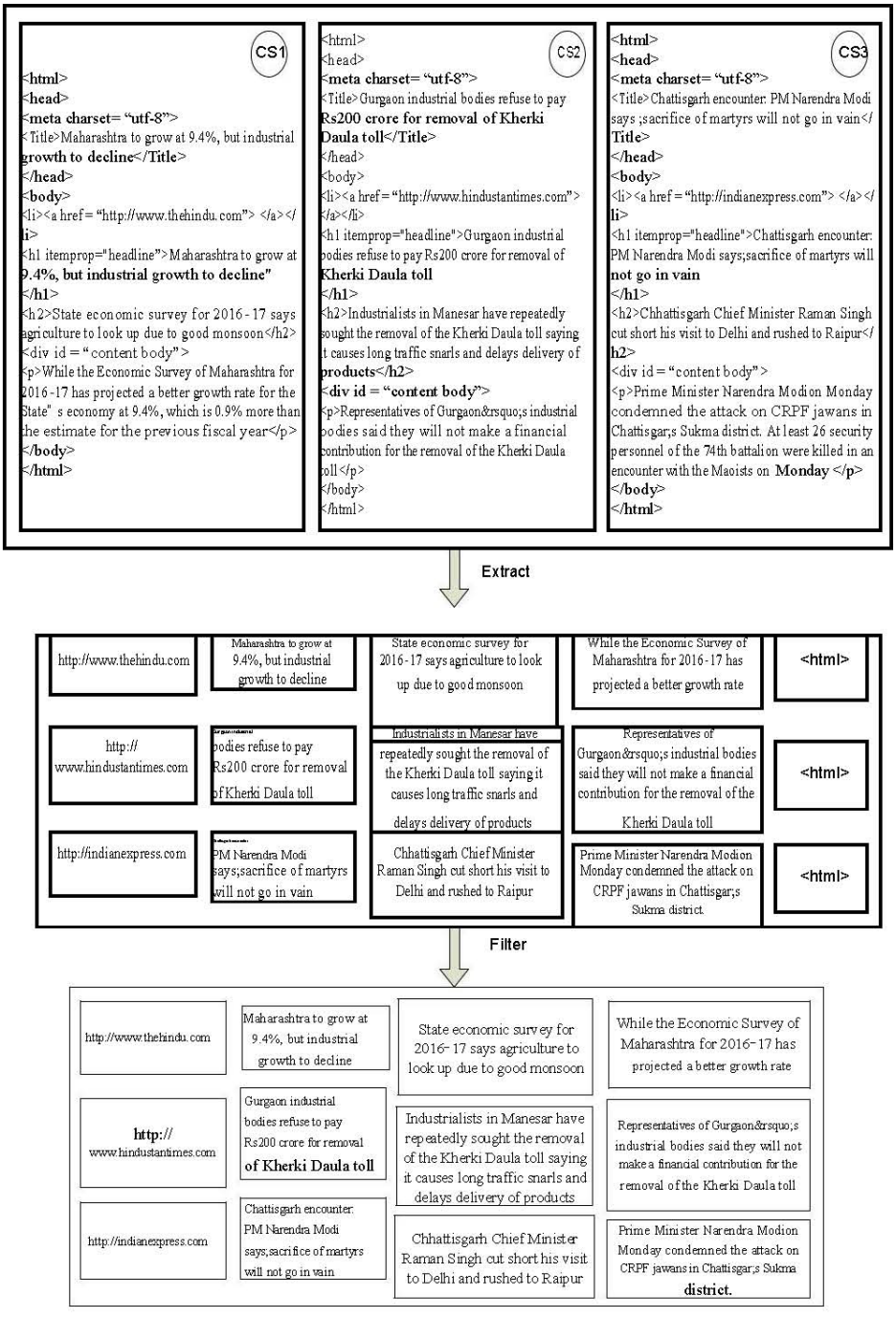
3.4 Filter algorithm

After extraction and pattern matching filtering algorithm is applied. Algorithm 4 presents the filter algorithm. Extraction algorithm returns a list of ContentSet that are supposed to contain the variable information in the initial ContentSet. In this algorithm, the main loop at lines 3-7 iterates over the list of input ContentSet and simply removes those in which inconsistency in the results. Let CS be the ContentSet, and SCS is the set of ContentSet. The result of filter algorithm is shown in Figure 7.

Algorithm 4 Filter

```
1  Filter (SCS: List ⟨ContentSet⟩): List ⟨ContentSet⟩
2      Result = ⟨⟩
3      Foreach CS in SCS do
4          If CS has inconsistency then
5              add CS to result
6          end
7      end
8  Return result
```

Figure 7 Filtering results after extraction (see online version for colours)



4 Dataset

The dataset used in our experiment contains a total of 500 news web documents. For experiments, we collected web news pages from ten news websites written in English. The sites are shown in Table 1. News web pages belong to different categories like business, cricket, India, tech, nation, science and environment, politics, world, entertainment, sports. Each category was randomly selected from Google search engine between December 2016 to March 2017. We downloaded 50 web pages from each web site.

Table 1 New websites

No.	*News website*	*URL*	*Category*
1	The Hindu	http://www.thehindu.com/	Business
2	The Times of India	http://timesofindia.indiatimes.com/	Cricket
3	NDTV	http://www.ndtv.com/	India
4	Hindustan Times	http://www.hindustantimes.com/	Tech
5	Indian Express	http://www.indianexpress.com/	Nation
6	Zee News	http://www.zeenews.com/	Science and environment
7	News18	http://www.news18.com/	Politics
8	The Pioneer	http://www.dailypioneer.com/	world
9	Deccan Harald	http://www.deccanherald.com/	Entertainment
10	The Asian Age	http://www.asianage.com/	Sports

Table 2 Common errors

S. no.	*Error*
1	Error: <!DOCTYPE> is missing
2	Error: <s6> is not recognized
3	Error: <j> is not recognized
4	Error: <bw> is not recognized
5	Error: <zs> is not recognized
6	Error: <z> is not recognized
7	Error: <m> is not recognized
8	Error: <v> is not recognized
9	Warning: replacing invalid character code 131
10	Warning: discarding invalid character code 143
11	Warning: unescaped& which should be written as &amp
12	Warning: unescaped& or unknown entity “&:p”
13	Warning: unescaped& or unknown entity “&X”
14	Warning: <p> unexpected or duplicate quote mark
15	Warning: <p> missing „>” for end of tag
16	Warning: <j> missing „>” for end of tag
17	Warning: discarding unexpected <j>
18	Warning: unescaped& or unknown entity “&qT”

Table 2 Common errors (continued)

S. no.	*Error*
19	Warning: unescaped& or unknown entity "&Us"
20	Warning: discarding unexpected <bw>
21	Warning: unescaped& or unknown entity "&oX"
22	Warning: <I> attribute "4" lacks value
23	Warning: <I> missing „>" for end of tag
24	Warning: discarding unexpected <I>

News web documents preprocess by fixing their HTML code by HTML tidy (Raggett, 1998). It fixes web documents doctype declarations, adds missing end tags, and reports on unknown attributes if essential. Our dataset were gathered from real world news websites, they usually contained errors in their HTML code. Table 2 shows the results we have gathered regarding a subset of common HTML errors that are reported by HTML tidy. The full report is too large to reproduce in the paper. Our only purpose to use HTML tidy is to make it clear that we have dealt with actual documents.

5 Experiment and results

In this section we present the results of the experiments we have carried out to compare our approach to other techniques in the literature. We describe the dataset used in our experimental study in section 4. We compare our approach to other approaches which are commonly used in extracting HTML pages in the literature. We performed our experiments on a machine that was equipped with an Intel Core i3 processor that run at 2.40 GHz, had 2 GB RAM, Windows 7 pro 64 bit.

We analyse the performance of our approach using the three parameters precision, recall and F_{1-} measure.

5.1 Other approaches

We compare our technique with other existing techniques.

ECON (Guo et al., 2010): it takes a collection of news web pages and use HTML parser to create DOM tree. There is a node that wraps the entire contents of news with its subtrees, such node is known as summary-node. ECON finds a snippet-node which is the descendent of the summary-node. When snippet-node is found, then backtracks it until a summary-node is found, by which firstly wrapped the part of the news content, then backtracks from the snippet-node until a summary-node is found, and the entire content of news can be extracted after removing noise from the summary-node.

CoreEx (Prasad and Paepcke, 2008): they extract the main article from news web pages by using DOM tree where every node in the tree represents the HTML node of a web page. They score every node based on two counts, textCnt and linkCnt which means the amount of text and number of links it contains. Their algorithm runs on 1120 news web pages.

Table 3 Comparison of all three approaches

Category	*Our Approach*			*ECON*			*CoreEx*		
	Precision	*Recall*	F_1*-measure*	*Precision*	*Recall*	F_1*-measure*	*Precision*	*Recall*	F_1*-measure*
Business	0.97	0.96	0.96	0.82	0.028	0.054	0.82	0.028	0.054
Cricket	0.99	0.99	0.99	0.94	0.91	0.931	0.94	0.91	0.931
India	0.95	0.94	0.94	0.92	0.88	0.89	0.92	0.88	0.89
Tech	0.96	0.96	0.96	0.88	0.67	0.67	0.88	0.67	0.67
Nation	0.99	0.99	0.99	0.83	0.83	0.78	0.83	0.83	0.78
Science and environment	0.93	0.82	0.82	0.76	0.77	0.77	0.76	0.77	0.77
Politics	0.98	0.88	0.87	0.67	0.67	0.67	0.67	0.67	0.67
World	0.95	0.96	0.87	0.91	0.88	0.86	0.91	0.88	0.86
Entertainment	0.96	0.94	0.94	0.73	0.68	0.68	0.73	0.68	0.68
Sports	0.92	0.87	0.87	0.64	0.76	0.76	0.64	0.76	0.76

The target of ECON and CoreEx is similar to our approach; both the approaches extract the main contents from news web pages. However, the underlying algorithms of both the approaches are subsequently different. According to Guo et al. (2010) and Prasad and Paepcke (2008) both ECON and CoreEx perform well, therefore we made a comparison between ECON, CoreEx and our approach.

5.2 Performance analysis

Experiment results are often evaluated by Precision, recall and F_1-measure. We first run ECON and CoreEx on the dataset in order to learn extraction rules, we then computed precision, recall and F_1-measure. For each website we recognised one type of template. Then we manually analyse URL's of these pages to identify repetitive patterns such as URL structure, home page, subsections, etc. A regular expression is written for each website to match the template with high interest to us. This way we handcrafted annotations for every web document in our dataset that is used to calculate precision, recall and accuracy of our proposed approach. We could find which extracted ContentSet was the closest to each annotation. For the validation of our approach we compared each extracted ContentSet to every annotation.

The precision of a given category of dataset is the fraction of web pages of its computed category that are also found in the corresponding annotated category of the dataset. The recall of a given category is the fraction of web pages from the corresponding annotated category of dataset that were extracted from the same annotated category. To calculate these measures, we assign two or more web pages to the same categories if and only if they are similar. A true positive (TP) decision assigns two structurally similar web pages to the same category; a true negative (TN) decision assigns two structurally different web pages to different categories. A false positive (FP) decision assigns two structurally unlike web pages to the same category. A false negative (FN) decision assigns two structurally similar web pages to different category. Then the precision and recall is calculated as follows:

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$f1 - measure = 2 * \frac{P * R}{P + R} \tag{3}$$

Table 3 shows the precision, recall and F-measure results of all three approaches ECON, CoreEx and our approach when run on all 500 news pages. The result of the extraction process is always a collection of ContentSet. Experimental results show that the performance of our approach is higher than the other two approaches.

6 Conclusions and future work

In this paper, we presented a news web page content extraction approach to automatically extract content from news web pages. It is based on the idea that web documents from different websites share tokens, and these tokens generate Tag tree. Experiments showed that our approach improves existing results in the literature (Guo et al., 2010; Prasad and Paepcke, 2008) for the problem of information extraction from news web pages and perform extraction with high precision and recall. Our approach applied to web news pages written in English. A news web page content extraction based on tag tree is proposed to efficiently extract meaningful information including records and data schema. In particular we have addressed the problem of finding and fetching news available on websites and extracting the relevant content. Through experimentation with ten news websites, we have demonstrated that our approach is highly effective for these tasks.

As future work, since the blog page and forum page have similar characteristics with news web pages, we will improve our approach to extract the content from blog and forum page.

References

Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M. and Etzioni, O. (2007) 'Open information extraction from the web', *20th International Joint Conference on Artificial Intelligence*, January, Vol. 7, pp.2670–2676.

Gibson, J., Wellner, B. and Lubar, S. (2007) 'Adaptive web-page content identification', in *Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management*, ACM, November, pp.105–112.

Guo, Y., Tang, H., Song, L., Wang, Y. and Ding, G. (2010) 'ECON: an approach to extract content from web news page', in *2010 12th International Asia-Web Conference (APWEB)*, *Pacific*, IEEE, April, pp.314–320.

Gupta, S., Kaiser, G.E., Grimm, P., Chiang, M.F. and Starren, J. (2005) 'Automating content extraction of html documents', *World Wide Web*, Vol. 8, No. 2, pp.179–224.

Ji, X., Zeng, J., Zhang, S. and Wu, C. (2010) 'Tag tree template for web information and schema extraction', *Expert systems with Applications*, Vol. 37, No. 12, pp.8492–8498.

Kaddu, M.R. and Kulkarni, R.B. (2016) 'To extract informative content from online web pages by using hybrid approach', in *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, IEEE, March, pp.972–977.

Knoblock, C.A., Lerman, K., Minton, S. and Muslea, I. (2003) 'Accurately and reliably extracting data from the web: a machine learning approach', in *Intelligent exploration of the web*, Physica-Verlag HD, pp.275–287.

Lin, S.H. and Ho, J.M. (2002) 'Discovering informative content blocks from web documents', in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, July, pp.588–593.

Louvan, S. (2009) *Extracting the Main Content from HTML Documents* [online] http://wwwis.win.tue.nl/bnaic2009/papers/bnaic2009_paper_113.pdf.

Muslea, I., Minton, S. and Knoblock, C. (1999) 'A hierarchical approach to wrapper induction', in *Proceedings of the Third Annual Conference on Autonomous Agents*, ACM, April, pp.190–197.

Pettersson, E., Lindström, J., Jacobsson, B. and Fiebranz, R. (2016) 'HistSearch-implementation and evaluation of a web-based tool for automatic information extraction from historical text', in *3rd HistoInformaticsWorkshop*, Krakow, Poland, 11 July.

Prasad, J. and Paepcke, A. (2008) 'Coreex: content extraction from online news articles', in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ACM, October, pp.1391–1392.

Raggett, D. (1998) 'Clean up your web pages with HP's HTML tidy', *Computer Networks and ISDN Systems*, Vol. 30, No. 1, pp.730–732.

Reis, D.C., Golgher, P.B., Silva, A.S. and Laender, A. (2004) 'Automatic web news extraction using tree edit distance', in *Proceedings of the 13th international conference on World Wide Web*, ACM, May, pp.502–511.

Sleiman, H.A. and Corchuelo, R. (2013) 'Tex: an efficient and effective unsupervised web information extractor', *Knowledge-Based Systems*, Vol. 39, pp.109–123.

Ziegler, C.N. and Skubacz, M. (2007) 'Content extraction from news pages using particle swarm optimization on linguistic and structural features', in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE Computer Society, November, pp.242–249.