International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 3 No. 1 May 2013, pp. 121-129

http://www.issr-journals.org/ijias/

Literature Review of Automatic Multiple Documents Text Summarization

Md. Majharul Haque[1], Suraiya Pervin[1], and Zerina Begum[2]

[1]Department of Computer Science & Engineering,
University of Dhaka,
Dhaka, Bangladesh

[2]Institute of Information Technology,
University of Dhaka,
Dhaka, Bangladesh

ABSTRACT: For the blessing of World Wide Web, the corpus of online information is gigantic in its volume. Search engines have been developed such as Google, AltaVista, Yahoo, etc., to retrieve specific information from this huge amount of data. But the outcome of search engine is unable to provide expected result as the quantity of information is increasing enormously day by day and the findings are abundant. So, the automatic text summarization is demanded for salient information retrieval. Automatic text summarization is a system of summarizing text by computer where a text is given to the computer as input and the output is a shorter and less redundant form of the original text. An informative précis is very much helpful in our daily life to save valuable time. Research was first started naively on single document abridgement but recently information is found from various sources about a single topic in different website, journal, newspaper, text book, etc., for which multi-document summarization is required. In this paper, automatic multiple documents text summarization task is addressed and different procedure of various researchers are discussed. Various techniques are compared here that have done for multi-document summarization. Some promising approaches are indicated here and particular concentration is dedicated to describe different methods from raw level to similar like human experts, so that in future one can get significant instruction for further analysis.

KEYWORDS: World Wide Web, search engine, information retrieval, document abridgement, human expert.

1 INTRODUCTION

The narration of automatic i.e. computerized abstraction began 60 years ago, as implementation of automatic text summarizer is often cited in the oldest publication in 1958 by H. P. Luhn [1]. The goal of automatic text summarization is to condense the given text to its essential contents, based upon user's choice of brevity. In this system, the summary is generated by machine to draw the most significant information in a shorter form of the source text, while still keeping its principal semantic content and helps the user to quickly understand large volumes of information. On the basis of methodology or techniques that are used for summarization, approaches can be divided into two broad groups – extraction and abstraction. Reformulation of contents is done while abstraction and the important sentences of original document are picked up in extraction. Extraction needs no background knowledge and this is domain independent, where abstraction is domain dependent in nature and requires human knowledge and is specific goal oriented [2]. Summarization task can be classified into two types [3]: 1) single document text summarization, 2) multi-document text summarization. After 2002, the single-document summarization task was approximately dropped [4]. In multi-document summarization, several key points are involved, such as reducing each document, incorporating all document's significant idea, compare the ideas found from each, ordering sentences come from different sources keeping the logical and grammatical structure right.

Corresponding Author: Md. Majharul Haque (mazharul_13@yahoo.com)

A range of procedures that employee document abstraction, such as neural networks, semantic graphs, fuzzy logic etc. are incorporated on the study on finding significant portion of text. The objective of this paper is to present a comprehensive literature review on automatic multi-document summarization using natural language processing and explore the trends of passage abstraction.

The rest of the paper is organized as follows. Section 2 briefly explains multi-document summarization. Section 3 presents a comprehensive literature review about different procedures on automatic multiple documents summarization. Section 4 turns conclusion with a brief about this paper.

2 MULTI-DOCUMENT TEXT SUMMARIZATION

Simply, multi-document text summarization means to retrieve salient information about a topic from various sources. Given a set of documents D = (D1, D2,…,Dn) on a topic T, the task of multi-document summarization is to identify a set of model units (S1,S2,…,Sn). The model units can be sentences, phrases or some generated semantically correct language units carrying some useful information. Then significant sentences are extracted from each model units and re-organized them to get multi-documents' summary. Process flow of multi-document summarization can be depicted as fig. 1.

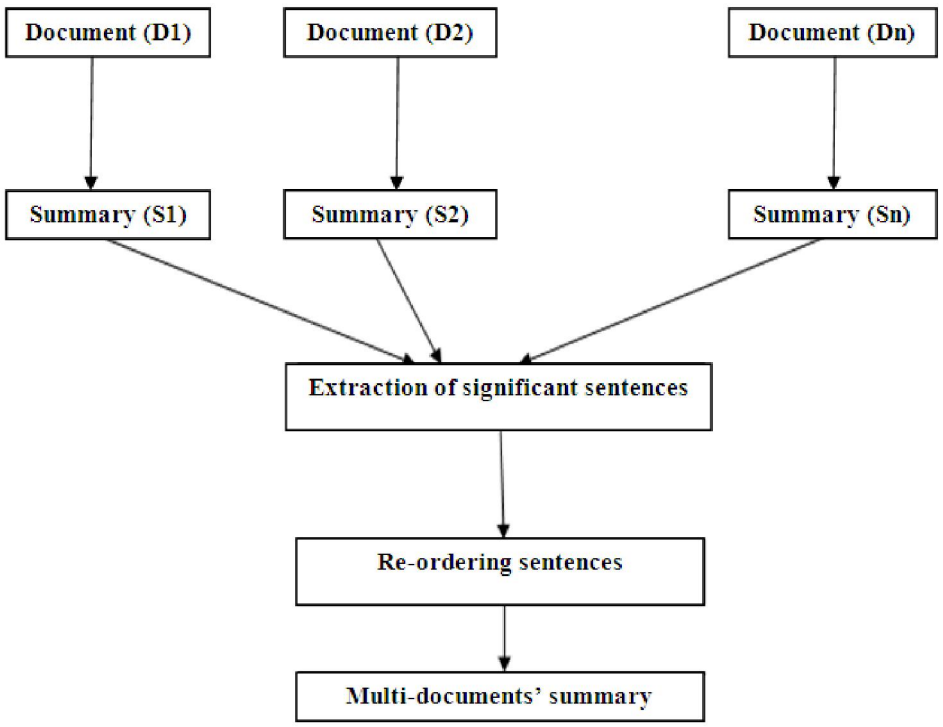

Fig. 1. Process flow of multiple-documents text summarization

3 REVIEW ON AUTOMATIC MULTIPLE DOCUMENTS TEXT SUMMARIZATION

Research on single document summarization has turned into its golden age and much more complicated to the era of multi-document summarization. Various methods are available for retrieval of information from diversified source. Many techniques have been developed from the very beginning of the thinking of multiple documents summarization to this time of modern science. In this paper, the discussed methods are grouped into five categories.

3.1 TERM FREQUENCY BASED METHOD

Salton [5] in 1989 introduced TF-IDF or term-frequency inverse-document-frequency model, where the score of a term in the document is the ratio of the quantity of terms in that document to the frequency of the quantity of documents containing that terms. The significance of term evaluation is given by the principle TFI X IDFI, where TFI is the frequency of term I in the document and IDFI is the inverted frequency of documents in which that term occurs. By computing pertinence of terms in the sentence, consequently sentences can be scored for illustration.

Jun'ichi Fukumoto [6] in 2004 offered a multi-document summarization technique that applied very simple strategy to generate abstract using TF/IDF based sentence extraction for single document summarization and use of single document summarization for multi-document. Their system automatically classifies a document set into three types using information of high frequency nouns and named entity: (a) one topic type, (b) multi-topic type, and (c) others. In the first type, the second document shows additional information or subsequent event of the first document, and so on for the following documents. In the second type, documents describe the same event type such as a set of traffic accidents. The third type of documents is related each other but not classified into the first two types. For summarization, at first sentences are extracted from each document based on TF/IDF, sentence position and weighing using intention type, such as "request", "obligation" and "necessary" etc., of a sentence. In the second step unnecessary parts of sentences are eliminated. Then extracted sentences are sorted in the original order in a document to generate condensed version of each single document and send them for document set type classification. After that all the extracted sentences are segmented into clauses and removed the repeated clauses and the rest of the clauses are sorted for generating expected summary. The task of document set type classification is a commendable effort in this research, but the mechanism used here for summarization is mostly based on single document abstraction.

You Ouyang et al [7] in 2009 introduced a novel hierarchical summarization approach which is able to integrate a range of objectives of multi-document abstraction. Human summarization concept is depicted here in such a way that man may start with finding the core topic in a document set and write something about this core topic. Next he may go to find sub-topic and sub-sub-topics and so on. Motivated by this experience, a hierarchical approach is offered here to mimic the behavior of human summarizer. The procedure includes two phases such as:

(i) Word hierarchical representation: Before constructing the hierarchical representation unnecessary concepts are removed from the document set where concepts are represented as terms of words. Two types of words are selected, i.e. query-relevance and topic-specificity and the identified keywords are sorted by their frequency. After that point-wise mutual information (PMI), is a measure of association used in information theory, is used to identify the subsumption between words and high PMI is regarded as significant. Using the identified relations, a top-down tree is constructed.

(ii) Summarization based on hierarchical representation: In this step sentences are selected through an iterative algorithm which follows a general to specific order. Words that are in the top level of the tree are regarded as the core concept. The algorithm moves to down level words through the subsumption relations between the words and new sentences are added except redundant sentences until the whole summary is generated.

Vikrant Gupta et al [8] in 2012 presented a new statistical approach to automatic summarization based on the Kernel of the source text called KernelSum (KERNEL SUMMarizer). Using simple statistical measures, Kernel is identified as the most significant passage of the source text. It serves as the guideline to choose the other sentences for summary. The procedure proposed here is composed with the following functional components: i) Text pre-processor works on converting the HTML or Word Documents to plain text, ii) Sentence separator divide the sentences based on some rules like ending point such as a dot and a space etc. iii) Word separator detaches the words through some criteria like a space, iv) Stop-words eliminator eradicates the regular English words like 'a, an, the, of form…', v) Word-frequency calculator computes the number of times a word appears in the document after removing stop-words, vi) Scoring algorithm estimate the score of each sentence by using the TF-ISF (Term Frequency - Inverse Sentence Frequency), vii) Ranking algorithm counts rank of every sentence according to the scores, location, length, heading sentence etc. viii) Summarizing part picked the sentences from the ranked list and concatenated to produce the expected brief of the input document. Final extract have been evaluated under the light of Kernel preservation and textuality and found 90% of the extracts have been judged to totally or partially preserve the gist, textuality was also highly graded: 85% of them were totally or partially coherent and cohesive.

3.2 GRAPH BASED METHOD

Inderjeet Mani et al [9] in 1997 represented topic through a set of entry nodes in the graph, along with edges corresponding to the semantic relations between items. The algorithm used here applies a spreading activation technique to discover nodes related to the core theme. The nodes whose meanings are equivalent to topic terms are treated as entry points into the graph and called activating node. Weight of nodes is an exponentially decaying function of activating node's weight and the distance between nodes. Weight of a neighbor node is calculated as a function of link weight and activating node weight. Consecutively the method finds neighbor of starting nodes and accumulate the activating nodes to the output until getting threshold number of output nodes.

In 2004 Rada Mihalcea et al [10] proposed an algorithm named TextRank using graph based method in the ground of natural language processing. A vertex is added for each sentence in the text to construct a graph. Link between vertices are set up using sentence similarity relation. This relation is based on content overlapping by which a score is generated for each vertex. After applying the iterative procedure consequently vertices are sorted by their scores, and then top scored sentences are chosen to construct abstract.

Junlin Zhang et al [11] in 2005 stated that multi-document extractive summarization depends on the notion of sentence centrality to recognize the most significant sentences in a document. A new approach under the hub-authority framework has been introduced here that unites the text content with some cues such as "cue phrase", "sentence length" and "first sentence" and investigates the sub-topics in the multi-documents by conveying the features of these sub-topics into graph-based sentence ranking algorithms. Old graph-based method is developed here with two essential different points: (i) unites the text content with some characteristics such as cue phrase, length of sentences and position. (ii) discovers the sub-topics with graph-based sentence ranking algorithms. Then the summary is generated according to the sentence ranking score of all sentences. The provided method was evaluated on DUC 2004 data and proved that the design of combining the exterior and interior features under the Hub/Authority framework is an effective graph-ranking schema in multi-document generic text abstraction.

Xiaojun Wan [12] in 2008 explored a graph-based ranking algorithm for multi-document summarization under the assumption that all the sentences are indistinguishable. Document impact on summarization performance is invented here with document-based graph model to incorporate the document-level information and the sentence-to-document relationship into the graph-based ranking process. Basic graph-based model is essentially a way of deciding the importance of a vertex within a graph based on global information recursively drawn from a one-layer link graph of sentences. The document-based graph model is integrated here to examine the document impact by exploring document importance and the sentence-to-document correlation into the sentence ranking process. This is a two-link graph including both sentences and documents. It is assumed that the sentences which belong to an important document, highly correlated with the document, will be more likely to be chosen into the summary.

Kokil Jaidka et al [13] in 2010 invented a novel summarization technique to generate literature review of research paper that mimics the characteristics of human literature reviews. An analysis has been carried out here to understand the human strategies of information selection and recapitulation. Some significant questions were thought before designing the procedure such as: i) where do researchers select information from? ii) what type of information do they select? iii) how do they fulfill the functions of a literature review? The novel approaches in this system would mainly be in the information selection and integration stage to select information from different semantic levels, and the rhetorical function implementation stage where the literature review will be drafted. In this proposed procedure three types of discourse structure are defined. For sentence-level, XML schema is constructed to define the valid XML document structure used to represent the structure of a literature review, including the expected elements and their hierarchical relationships. For clause-level and intra-clause-level, a graphical representation of rhetorical relations is represented as a tree structure between the constituent clauses of text. A number of strategies applied to select salient parts from this XML or graphical tree structure to produce a comparative literature review, such as: a) Correlation between the candidate topic and source content, b) Semantic similarity measures, c) Relative information gain ratios of information with respect to the surrounding text.

3.3 TIME BASED METHOD

McKeown et al [14] in 1995 presented a natural language processing system that summarizes a series of news articles on the same event using empirical analysis. Length of summaries varies on the basis of the available resources of text. The proposed system here named SUMMONS (SUMMarizing Online NewS articles) that summarize full text input using templates formed by the message understanding systems developed under the ARPA human language technology program [15]. Their research focused on techniques to summarize how the trends of an event changes over time, using various points of view over the same event or series of events. Input to SUMMONS is a set of templates, where each template represents the information extracted from one or more sources by a message understanding system. It first groups messages jointly, identifies commonalities between them, and notes how the discourse influences wording by setting realization flags. The departure point is in the stage of identifying what information to include and how to group it together.

Xiaojun Wan et al [16] in 2007 unveiled TimedTextRank algorithm as an enhancement of graph based ranking process namely TextRank for multi document summarization and incorporated a new temporal dimension. A proclamation has been made that for an evolving topic, recent documents are usually more important than earlier documents because of the

availability of novel information. The TextRank procedure makes use of the relationships between sentences and chooses sentences according to the "vote" or "recommendations" from their neighboring sentences, which is similar to PageRank and HITS. An affinity graph is generated at first to reveal the relationships among all sentences in the document set. Now the vote is casted for each node in a way that the votes cast from new documents are attached more importance than the votes cast from the sentences in old documents. By this way the informativeness score is calculated to select sentences for generating summary.

3.4 SENTENCE CO-RELATION BASED METHOD

S. Hariharan et al [17] in 2012 proposed enhancements on two graphical methods namely- LexRank (threshold) and LexRank (continuous) offered by Erkan and Radev [18]. LexRank and Continuous LexRank techniques are developed based on modification of the most popular page ranking algorithms designed for web link analysis. A link between two sentences is considered as a vote cast from one sentence to the other sentence. The score of a sentence is determined by the votes that are casted for it and the scores of the sentences casting these votes. A document can be considered as a network of sentences those are associated with each other. Cosine similarity has been used to discover similarity between two couple of sentences and to assess the relevance between sentences. Proposed enhancements in this paper are discounting technique and position weight factor. Discounting method envisages that once a sentence is selected then the next sentence is selected based on the contributions made by the remaining n-1 sentences only. So, the chance for repetition of information in the succeeding sentences is minimized, and the summary will be cohesive and meaningful. In the graph based approach, importance to position of the sentence can be given in a way by giving preference to sentences that occurs earlier out of two documents considered. For instance, first sentence in a document of 5 sentences will get a weight of 1/5 = .20 and in a document of 10 sentences will get a weight of 1/10 = .10. Now the node will be extracted from graph based on casted votes, scores, position, weight, etc. to get the abstract.

Tiedan Zhu et al [19] in 2012 in their paper proposed the logical closeness criterion to measure the similarity between two sentences through which extracted sentences for summarization can be chronologically ordered. In multi-document summarization, sentences are selected from various documents differ with single document summarization. So a strategy to arrange the order of sentences is demanded. This publication also gave a brief review about the others work on sentence ordering and offered an improved procedure. This research emphasized on logical-closeness rather than topical-closeness which is based on synonymy and not strong enough to measure the coherence of sentences. To assess logical-closeness following techniques are applied,

(i) Notation Definition: the arrow '→' and the sentence-chain. For two sentences A and B, a notation is defined A→B to represent that A and B are adjacent where A precedes B. Multiple sentences are represented as chain with arrow.

(ii) Definition of Logical-closeness: logical-closeness means the closeness in meanings. Sometime two sentences have no topical-closeness but coherent in sense.

(iii) Measure of Logical-closeness: If sentence A is similar with the adjacent sentences of B and vice versa then sentences A and B are coherent. By this way coherency with each other is calculated.

Finally more adjacent sentences are picked up in a chain to produce the logical summary.

3.5 CLUSTERING BASED METHOD

Jade Goldstein et al [20] in 2000 presented a method for text extraction approach to multi-document summarization that builds on single-document summarization methods by using supplementary available information about the document set and relationships between the documents. Here they identified four minimum requirements for multi-document summarization: (a) clustering- the ability to cluster similar documents and passages to find related information, (b) coverage- the ability to find and extract the main points across documents, (c) anti redundancy- the ability to minimize redundancy between passages in the summary, (d) summary cohesion criteria- the ability to combine text passages in a useful manner for the reader. The proposed procedure emphasized on "relevant novelty" which is a metric for minimizing redundancy and maximizing both relevance and diversity. The method works as follows: (i) Segment the documents into passages, and index them, (ii) Identify the passages relevant to the query using cosine similarity with a threshold below which the passages are discarded, (iii) Using "relevant novelty" metric, depending on the desired length of summary, select a number of passages, (iv) Resemble the selected passages into a summary document.

Judith D. Schlesinger et al [21] in 2008 proposed a multi-document summarization technique that combines Clustering, Linguistics and Statistics for Summarization and named it CLASSY. It uses linguistic trimming and statistical methods to

produce generic summaries for both single and clusters of documents. CLASSY has cut a good figure in the Document Understanding Conference (DUC) evaluations and Multi lingual (English and machine translated/original version of Arabic document) Summarization Evaluations (MSE). The proposed method used trimming rules to shrink sentences, identify sentences and organize the chosen sentences for the final summary. Here the thought was to design a multi-lingual summarization technique. CLASSY structural design made up of five steps: preparation of raw texts, trimming of sentences, scoring, redundancy elimination and sentence organizing. This method can also be used for machine translated edition of Arabic document as well as English document. The trimming method is truly dependent on language and the quality of summarization very much depends on the translation quality of machine.

Xiaojun Wan et al [22] in 2008 presented a summarization procedure using cluster-based link analysis. This paper described about Markov Random Walk model which exploited for multi-document summarization by making use of the link relationships between sentences in the document set. Two models were proposed here to incorporate the cluster-level information into the process of sentence ranking. The first model is the Cluster-based Conditional Markov Random Walk Model (ClusterCMRW), which incorporates the cluster-level information into the link graph. The second model is the Cluster-based HITS Model (ClusterHITS), which considers the clusters and sentences as hubs and authorities in the HITS algorithm. Both models are based on link analysis techniques. The overall summarization framework consists of the following three steps:

(i) Theme cluster detection: By using clustering algorithm this method detect theme cluster from the document set.

(ii) Sentence score computation: This step aims to compute the saliency scores of the sentences in the document set by using either the ClusterCMRW model or the ClusterHITS model to incorporate the cluster-level information.

(iii) Summary extraction: In the final step, redundancy is removed and high scored sentences are chosen as summary sentences.

Nitin Agarwal et al [23] in 2011 presented an unsupervised approach called SciSumm for multi-document summarization of scientific articles. Using the context of the co-citation in the source article, the system produces a query by which it can generate a summary in a query-oriented fashion. In this proposed method SciSumm has four principal modules that are central to the functionality of the system. i) TextTilling module: This module used to obtains tiles of text relevant to the citation context using TextTilling algorithm [24]. Those text tiles are used as the basic unit for summary. ii) Clustering module: Frequent Term based text clustering algorithm [25] is used to generate labeled clusters using the text tiles extracted from the co-cited paper. iii) Ranking module: The clusters are ordered according to relevance with respect to the generated query using ranking module. iv) Summary presentation module: This module is used to display the ranked clusters obtained from the ranking module.

4 COMPARISON AMONG THE TECHNIQUES

At a glance comparison among the techniques of multi document text summarization has been shown in table 1:

Table 1. Comparison Among the Techniques of Multiple documents Text Summarization

#	Researcher(s), Year, Reference	Category	Basis of procedure
1	G. Salton, 1989, [5]	Term frequency based method	The significance of term evaluation is given by the principle TFI X IDFI, where TFI is the frequency of term I in the document and IDFI is the inverted frequency of documents in which that term occurs.
2	Jun'ichi Fukumoto, 2004, [6]	Term frequency based method	This method applied very simple strategy to generate abstract using TF/IDF based sentence extraction for single document summarization and use of single document summarization for multi-document.
3	You Ouyang, 2009, [7]	Term frequency based method	The procedure includes two phases such as: word hierarchical representation on the basis of most frequent terms in top of hierarchy and summarization based on hierarchical representation.

4	Mr.Vikrant Gupta, 2012, [8]	Term frequency based method	Using simple statistical measures, Kernel is identified as the most significant passage of the source text that contains most frequent terms. It serves as the guideline to choose the other sentences for summary.
5	Inderjeet Mani, 1997, [9]	Graph based method	The algorithm used here applies a spreading activation technique to discover nodes related to the core theme. Consecutively the method finds neighbor of starting nodes and accumulate the activating nodes to the output.
6	Rada Mihalcea, 2004, [10]	Graph based method	A vertex is added for each sentence and link between vertices are set up using sentence similarity relation. Then top scored sentences are chosen to construct abstract.
7	Junlin Zhang, 2005, [11]	Graph based method	A new approach under the hub-authority framework has been introduced here that unites the text content with some cues and investigates the sub-topics into graph-based sentence ranking algorithm for generating expected output.
8	Xiaojun Wan, 2008, [12]	Graph based method	This is a two-link graph including both sentences and documents. It is assumed that the sentences which belong to an important document, highly correlated with the document, will be more likely to be chosen into the summary.
9	Kokil Jaidka, 2010, [13]	Graph based method	XML schema is constructed to define the valid XML document structure used to represent the structure of a literature review. Then a number of strategies applied to select salient parts from this XML or graphical tree structure to produce a comparative literature review.
10	Kathleen McKeown, 1995, [14]	Time based method	This method focused on techniques to summarize how the trends of an event changes over time, using various points of view over the same event or series of events.
11	Xiaojun Wan, 2007, [16]	Time based method	Here the enhancement of TextRank is unveiled named TimedTextRank with incorporating time dimension. This is based on the proclamation that for an evolving topic, recent documents are usually more important than earlier documents.
12	Shanmugasundaram Hariharan, 2012, [17]	Sentence co-relation based method	A link between two sentences is considered as a vote cast from one sentence to the other sentence. Sentences will be extracted based on casted votes, scores, position etc. to get the abstract.
13	Tiedan Zhu, 2012, [19]	Sentence co-relation based method	This research emphasized on logical-closeness rather than topical-closeness which is based on synonymy and not strong enough to measure the coherence of sentences.
14	Jade Goldstein, 2000, [20]	Clustering based method	Using clustering, coverage, anti redundancy and summery cohesion criteria the proposed procedure emphasized on “relevant novelty” which is a metric for minimizing redundancy and maximizing both relevance and diversity.

15	Judith D. Schlesinger, 2008, [21]	Clustering based method	This method combines Clustering, Linguistics and Statistics for Summarization and named it CLASSY. Structural design made up of five steps: preparation of raw texts, trimming of sentences, scoring, redundancy elimination and sentence organizing.
16	Xiaojun Wan, 2008, [22]	Clustering based method	Two models were proposed here in the process of sentence ranking. One is to incorporate the cluster-level information into the link graph. Second is to consider the clusters and sentences as hubs and authorities in the HITS algorithm for scoring sentences.
17	Nitin Agarwal, 2011, [23]	Clustering based method	This technique generates a summary in a query-oriented fashion with an unsupervised approach called SciSumm. Here the proposed method has four principal modules: text tilling, clustering, ranking and summery presentation.

5 CONCLUSION

In this paper, concepts of multiple documents text summarization are reviewed that categorize different approaches in this ground. This literature review explore the recent trend in summarization system that comes from novice procedure to this time of computer, where natural language processing is used to generate the summary resemble with human expert. Almost all the techniques found for summarization presumed that the documents of correlated topic will be submitted for abstraction. Though Fukumoto J. in 2004 classified given documents into three types, in third type it was assumed that all the documents have association with each other [6]. There is hardly any research found yet to categorize the presented documents with similarity measurement before summarizing. We have faith that the study on multi-document summarization system is a productive region for further research. Around 17 papers have been discussed here and various key topics from other historical publication relevant with text summarization have been analyzed here from 1988 to 2012. There exist some other techniques similar with those described in this paper, the discussion of which has not been included here as it will be a large corpus. But it is expected that any researchers can get help from this literature review for better understanding of different types of procedure on multi-document summarization. Anyone can also get direction for better perception of the diversified sorts of abstraction, which will help to construct new procedure for next generation.

REFERENCES

[1] Hans P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159-165, 1958.

[2] B. Cretu, Z. Chen, T. Uchimoto and K. Miya, "Automatic Summarization Based on Sentence Extraction: A Statistical Approach," *International Journal of Applied Electromagnetics and Mechanics*, IOS Press, vol. 13, no. 1-4, pp. 19-23, 2002.

[3] S. Suneetha, "Automatic Text Summarization: The Current State of the art," *International Journal of Science and Advanced Technology* (ISSN 2221-8386), vol. 1, no. 9, pp. 283-293, 2011.

[4] Krysta M. Svore, Lucy Vanderwende and Christopher J.C. Burges, "Enhancing Single-document Summarization by Combining RankNet and Third-party Sources," Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 448–457, 2007.

[5] G. Salton, "Automatic Text Processing: the transformation, analysis, and retrieval of information by computer," Addison-Wesley Publishing Company, USA, 1989.

[6] Jun'ichi Fukumoto, "Multi-Document Summarization Using Document Set Type Classification," Proceedings of NTCIR- 4, Tokyo, pp. 412-416, 2004.

[7] You Ouyang, Wenji Li and Qin Lu, "An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation," Proceedings of the ACL-IJCNLP Conference Short Papers, Suntec, Singapore, pp. 113–116, 2009.

[8] Mr.Vikrant Gupta, Ms Priya Chauhan, Dr. Sohan Garg, Mrs. Anita Borude and Prof. Shobha Krishnan, "An Statistical Tool for Multi-Document Summarization," *International Journal of Scientific and Research Publications* (ISSN 2250-3153), vol. 2, issue 5, 2012.

[9] Inderjeet Mani and Eric Bloedorn, "Multi-document summarization by graph search and matching," AAAI/IAAI, vol. cmp-lg/9712004, pp. 622-628, 1997.
[10] Rada Mihalcea and Paul Tarau, "Text-rank: Bringing Order into Texts," Proceeding of the Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004.
[11] Junlin Zhang, Le Sun and Quan Zhou, "A Cue-Based Hub-Authority Approach for Multi-Document Text Summarization," Proceedings of NLP-KE'05, IEEE, pp. 642-645, 2005.
[12] Xiaojun Wan, "An Exploration of Document Impact on Graph-Based Multi-Document Summarization," Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, pp. 755-762. 2008.
[13] Kokil Jaidka, "Multidocument Summarization of Information Science Research Papers," Bulletin of IEEE Technical Committee on Digital Libraries: JCDL/ICADL Doctoral Consortium Issue, vol. 6, issue 2, 2010.
[14] Kathleen McKeown and Dragomir R. Radev, "Generating Summaries of Multiple News Articles," Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, Seattle, Washington, pp. 74–82, 1995.
[15] Message Understanding Conference (MUC). Proceedings of the Fourth Message Understanding Conference (MUC- 4). DARPA Software and Intelligent Systems Technology Office, 1992.
[16] Xiaojun Wan, "TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization," Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, pp. 867- 868, 2007.
[17] Shanmugasundaram Hariharan, Thirunavukarasu Ramkumar and Rengaramanujam Srinivasan, "Enhanced Graph Based Approach for Multi Document Summarization," *The International Arab Journal of Information Technology*, 2012. [Online] Available: www.ccis2k.org/iajit/PDF/vol.10,no.4/4460-11.pdf (March 11, 2013)
[18] G. Erkan and D. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457-479, 2004.
[19] Tiedan Zhu and Xinxin Zhao, "An Improved Approach to Sentence Ordering For Multi-document Summarization," IACSIT Hong Kong Conferences, IACSIT Press, Singapore, vol. 25, pp. 29-33, 2012.
[20] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, "Multi-Document Summarization by Sentence Extraction," ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, *New Jersey*, pp. 40-48, 2000.
[21] Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy, "Arabic/English Multi-document Summarization with CLASSY - The Past and the Future," Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, Haifa, Israel, pp. 568–581, 2008.
[22] Xiaojun Wan and Jianwu Yang, "Multi-Document Summarization Using Cluster-based Link Analysis," Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Singapore, pp. 299-306, 2008.
[23] Nitin Agarwal, Gvr Kiran, Ravi Shankar Reddy and Carolyn Penstein Ros´e, "Towards Multi-Document Summarization of Scientific Articles:Making Interesting Comparisons with SciSumm," Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, pp. 8–15, 2011.
[24] Marti A. Hearst, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages," Computational Linguistics, vol. 23, no. 1, pp. 33-64, 1997.
[25] Florian Beil, Martin Ester and Xiaowei Xu, "Frequent Term-Based Text Clustering," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada, pp. 436-442, 2002.