International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 10, October - 2013

Document Categorization using Data Mining Techniques

Ms.Sunita R. Patil, Student, NMIMS, Mumbai,

Dr.Mrs.Sunita M. Mahajan, Principal, ICS, MET, Bandra,

***Abstract*—Technical articles published worldwide have most of similarity in contents and repeated relevant information. Thus reading these related articles to get the recent research developments is time-consuming, unnecessary, irrelevant, cumbersome and impossible.**

To solve above problems an innovative system is developed which summarizes these articles using various Data Mining strategies. The system is named '*Optimized Summary System (OSS)*' consisting three models as 'Research Relevant Novel' (RRN) terms identification, category generation and document optimization through 'Maximal Marginal Relevance' (MMR).

***OSS* gives short condensed and accurate summarized categorical contents presenting innovative authentic information from multiple relevant technical articles.**

***Index Terms*—Optimization; summarization; RRN; MMR**

I. INTRODUCTION

The huge amount of technical information is published worldwide in the form of research articles every year. Reading these multiple domain specific articles one by one to get desired information is just time-consuming, sometimes unnecessary, irrelevant and impossible. The person who is referring these articles also needs to know advantages, drawbacks, overview of the evaluation methodologies and typical numerical results for future reference. Therefore, it is the need of the day to summarize these articles and only present the short, condensed, most relevant and accurate topic specific information.

We designed an innovative solution called '*Optimized Summary System (OSS)'* which provides single as well as multiple research articles' aggregation reducing redundancies. *OSS* helps reader to get each as well as multiple article's *research purposes, approaches, techniques & methodologies* used, *earlier research developments, continuation of existing research or novel ideas,* author's *own* and or *other researcher's work, results, discussions and outcomes.* The article summary is produced under multiple research relevant categories. These categories are nothing but the most relevant extracted sentences using research oriented terms, we named them as 'Research Relevant Novel (RRN) terms. The Maximal Marginal Relevance (MMR) criteria are used for optimization and reducing redundancies [10]. Various Data Mining Techniques are used for extraction and clustering of technical articles providing collective contents and allowing the reader to access their main topic of interest only.

II. LITERATURE REVIEW

Earliest text based summarization articles proposed few techniques like word and phrase frequency [2], position in the text [6], key phrases [3] and text processing [7]. These articles worked mostly on news related textual data [8, 9]. There are extractive *summarization* techniques available which rely on extraction of sentences. These extract sentences are revised by deleting or inserting terms or phrases [10].Whereas the *abstractive summarization* techniques focuses on the use of advanced language generation approaches. Authors [4, 10, 11] employed decision tree learning (C4.5), whereas others employed the Bayesian classifiers. The *discourse structure information* source is used to determine the type for each span and the relation between spans, and to organize the spans into a tree structure, using cue words and phrases, and lexical repetition [9]. The *Rhetorical information* is also used [12] to improve the performance of a summarizer based on lexical chains [13]. Moreover, [14] present a method of automatically generating templates for summarization. Research confirms that the *lead summary* contains most of the important information from newswire texts [12, 15, 16]. But 30% lead summaries do not perform very well [17]. The current summarization techniques use *key phrase extraction* [19] from scientific articles. More recently *citation texts* are useful in creating a summary [19]. But using it directly is not suitable [18]. *Natural Language Processing* (NLP) and recommendation systems also do not build a complete solution for different tasks. Tackling the individual domain specific components as per the user need is still an unsolved problem.

Our *Optimized Summary System* (*OSS*) using RRN terms and term based RRN categories expected by the user will be the best solution for getting summarized contents from scientific articles.

III. SYSTEM DESIGN

In scientific papers all the basic research oriented statements are almost covered in '*Abstract'*, '*Introduction'* and *'Conclusion'* sections. But the research oriented statements in these sections are almost repeated. This information repeatedly *covers* research goals, methods and techniques used, earlier research extensions or new ideas and results with outcomes. Thus we have selected only '*Abstract'*, and '*Introduction'* sections instead of whole document for optimized summarization. The *OSS* architecture is shown in Figure.1. It consists of three main modules as follows:

i. Research Relevant Novel (RRN) term identification and Categorization
ii. Maximal Marginal Relevance Metric Generation and
iii. Summary Generation using Data Mining Strategies

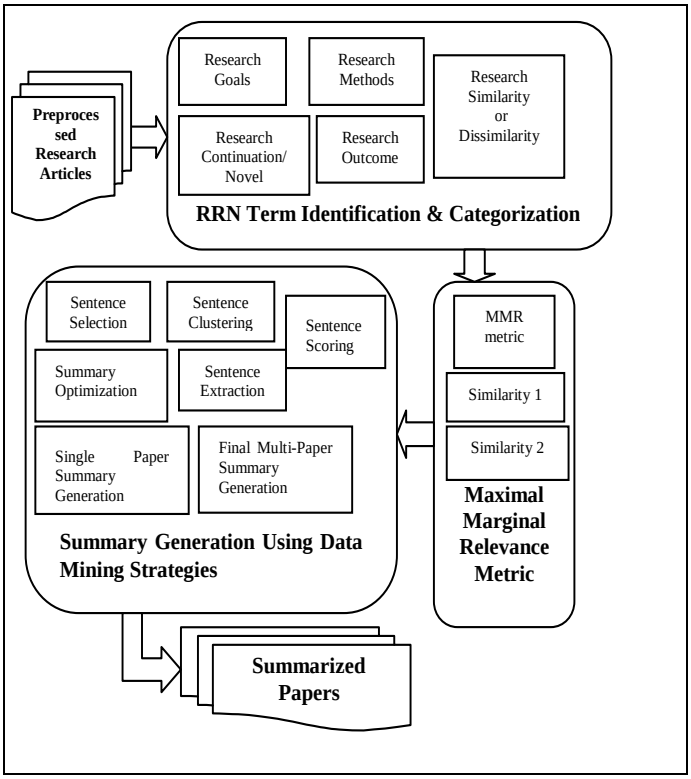

Fig 1. Optimized Summary System, (OSS) Architecture

The research articles to be summarized are first preprocessed using *sections segmentation* and *converted into plain text*. It removes all *stopwords* and *tokenizes* the article. The terms consisting research oriented words are identified as '*Research Relevant Novel* (RRN)' terms and reader expected *research categories* such as *Research Goals, Research Methods, Research Similarity or Dissimilarity, Research Continuation / Novel* and *Research Outcome* are generated using '*RRN Term Identification and Categorization'* module. Table I shows the RRN Categories generated as per their sentence roles.

TABLE I.
RRN Categories and their Sentence Roles.

RRN Category	***Sentence Role***
Research Goal	Sentences representing the purpose or aim or principle innovative idea of research under study for current paper;
Research Methods	Sentences representing the methods or approaches or ways used for the goal achievement;
Research Contrast/ Like	Sentences claiming authors own work contrast with others/ earlier work; showing limitations in others/ earlier work; direct comparing with others/ earlier work; research work of this kind never done before; sentences presenting similarity with others / earlier work;
Research Continuation	Sentences with research continuation of earlier/existing work;
Research Outcome	Sentences relating to result, conclusion, outcome; showing end product; Sentences stating evaluation of implementation;

The Maximal Marginal Relevance (MMR) metric [10] criterion reduces redundancy while maintaining maximum query relevance measuring relevance and novelty independently.

MMR metric is defined as,

$$MMR(P,C,Q,R,S) = Arg\ \max_{P_{ij}\ \epsilon\ R\setminus S}[\lambda * Sim_1(P_{ij},Q,C_{ij}) - (1-\lambda) * \max_{P_{nm}\ \epsilon\ S}(Sim_2(P_{ij},P_{nm},C,S))] \quad (1)$$

Where, $\boldsymbol{Sim_1}$ is the similarity metric for relevance ranking and $\boldsymbol{Sim_2}$ is the similarity metric for anti-redundancy as given below:

$\boldsymbol{Sim_1}$= (*cosine similarity* metric of sentence and query + *coverage* score for the sentence by whether the sentence is *in one or more clusters* and the *size* of the cluster +*information content* of the sentence by taking into account the *RRN terms).* (2)

$\boldsymbol{Sim_2}$= (cosine *similarity* metric of sentence and previously selected sentence + penalize sentences that are part of clusters from which other sentences have already been chosen + penalize documents from which *sentences* have *already been selected*) (3)

Data Mining Strategies are then used for category-wise relevant *Sentence Extraction* followed by *Sentence Clustering*, *Sentence Selection* and *Sentence Scoring* in order to assign an important score to each sentence using the modules respectively. Finally the summary is generated as single paper or multiple paper summaries using '*Single Paper'* and '*Multi-Paper Summary Generation'* module. To select summary sentences centroid sentence, TF*IDF and Term Frequency features are used where the *term frequency, tf* (*t, d*) is the occurrence count of a RRN term '*t*' in a research article document '*d*' with article collection *'D'*. The *inverse document frequency, IDF(t)* which is a measure of the general importance of the terms from research paper document collection and since IDF on its own is a relatively weak indicator of the term's importance and for this reason very often it is used in conjunction with the term frequency (TF) as $TF-IDF(t,d)$ which are given in the equations 4,5 and 6 respectively.

$tf(t_{i,j},d)$ = raw frequency of the RRN terms at position $_{i,j}$ in document *d*. (4)

$$IDF(t) = log(|D|/|\{d: t \in d\}| \quad (5)$$

$$TF-IDF(t,d) = tf(t,d) \times idf(t) \quad (6)$$

The selected sentences are then clustered into five RRN categories using cosine similarity between them. The score of each RRN term category is calculated by measuring term frequency of RRN terms i.e. the score of each term.

User can select number of sentences per research category to form individual paper summary. For final multi-paper summarization, the elected sentences are stored till the desired percentage for summarization is met for each single paper summary under RRN categories. In order to generate desired percentage of summary, a threshold is set as:

Threshold = ((Total sentences of first document) + (Total sentences of second document) + (Total sentences of third

document) +…+ (Total sentences of n^{th} document)) × (Desired percentage of summary). (7)

By comparing multiple papers' query topic, the reader can decide which document should be selected for further in depth reading.

IV. RESULTS

The standard formatted published research papers from IEEE Explore are selected for summarization and are made ready by preprocessing them. Figure. 2 to 8 shows the results of various *OSS* architecture modules for input document no. 1 and fig. 9 shows output of two input documents as research papers.

RRN Terms Identification
paper proposes, in this paper, we present, a new approach, it is very essential, conveying the same fact, concept, this poses, a significant impact, our HDS uses, to produce, the evaluation, we propose, this results in, our proposed, aims to deal with, in this work, we try to catch, we modularize, the basic computation, be interpreted, is designed, in such a way, the main challenge, this work, this method is, to analyze, results ,purpose

Figure 2. RRN Term Identification

Research Category Generation	***RRN Terms***
[Research Goals]	paper proposes, in this paper, we present, concept, this poses, it is very essential, we propose, our proposed, , aims to deal with, to analyze, the purpose
[Research Methods]	a new approach, in this work, we modularize, the basic computation, is designed, in such a way, this work, this method is, our HDS uses, to produce
[Research Continuation/Novel]	conveying the same fact, we try to catch
[Research Similarity/dissimilarity]	in the context of
[Research Outcome]	the evaluation, , this results in, be interpreted, the main challenge, results , a significant impact,

Figure 3. RRN Term Category Generation

Sentence Extraction
1. In this paper, we present a new approach that incorporates semantic information from a document, in the form of Hierarchical Document Signature (HDS), to measure semantic similarity between sentences. 2. Due to variability of expressions of natural language, it is very essential to exploit the semantic properties of a document to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different. 4. This poses a significant impact on many text mining applications performance where sentence-level judgment is involved. 5. Our HDS uses the natural hierarchy of the document and represents it in a modularized form of document level to sentence level, sentence to word level; aggregating similarity components at the lower levels and propagating them to the next higher level to produce the final similarity between sentences. 6.The evaluation of our HDS model has shown that it resembles the decision making process as done by human to a greater extent than different vector space models which only uses 'bag of words' concept. 7. In this paper, we propose an application of hierarchical document signature (HDS), extension of fuzzy signature that takes into account semantic structure of sentences to measure sentences similarity. 9. This results in "semantic loss" because semantic contextual senses of the sentences are discarded. 11. Our proposed method aims to deal with this issue by utilizing semantic similarity of constituent words in the sentences and then using that information to find the overall similarity between pairs of sentences using HDS structure. 13. Each imparts different information to the context of a sentence.

Figure 4. Sentence Extraction

Sentence Similarity Clustering	
[Research Goals]	1. In this paper, we present a new approach that incorporates semantic information from a document, in the form of Hierarchical Document Signature (HDS), to measure semantic similarity between sentences. 7. In this paper, we propose an application of hierarchical document signature (HDS), extension of fuzzy signature that takes into account semantic structure of sentences to measure sentences similarity. 11. Our proposed method aims to deal with this issue by utilizing semantic similarity of constituent words in the sentences and then using that information to find the overall similarity between pairs of sentences using HDS structure. 13. Each imparts different information to the context of a sentence.
[Research Methods]	4. This poses a significant impact on many text mining applications performance where sentence-level judgment is involved. 5. Our HDS uses the natural hierarchy of the document and represents it in a modularized form of document level to sentence level, sentence to word level; aggregating similarity components at the lower levels and propagating them to the next higher level to produce the final similarity between sentences.
[Research Continuation/Novel]	2. Due to variability of expressions of natural language, it is very essential to exploit the semantic properties of a document to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different.
[Research Similarity/Dissimilarity]	13. Each imparts different information to the context of a sentence.
[Research Outcome]	6.The evaluation of our HDS model has shown that it resembles the decision making process as done by human to a greater extent than different vector space models which only uses 'bag of words' concept. 9. This results in "semantic loss" because semantic contextual senses of the sentences are discarded.

Figure 5. Sentence Similarity Clustering

Sentence Scoring
[Research Goals]
1. In this paper, we present a new approach that incorporates semantic information from a document, in the form of Hierarchical Document Signature (HDS), to measure semantic similarity between sentences. **Score:04** 7. In this paper, we propose an application of hierarchical document signature (HDS), extension of fuzzy signature that takes into account semantic structure of sentences to measure sentences similarity. **Score:03** 11. Our proposed method aims to deal with this issue by utilizing semantic similarity of constituent words in the sentences and then using that information to find the overall similarity between pairs of sentences using HDS structure. **Score:03** 13. Each imparts different information to the context of a sentence. **Score:02**
[Research Methods]
4. This poses a significant impact on many text mining applications performance where sentence-level judgment is involved. **Score:03** 5. Our HDS uses the natural hierarchy of the document and represents it in a modularized form of document level to sentence level, sentence to word level; aggregating similarity components at the lower levels and propagating them to the next higher level to produce the final similarity between sentences. **Score:04**

[Research Continuation/Novel]

2. Due to variability of expressions of natural language, it is very essential to exploit the semantic properties of a document to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different.
Score:02

[Research Similarity/Dissimilarity]

13. Each imparts different information to the context of a sentence.
Score:01

[Research Outcome]

6.The evaluation of our HDS model has shown that it resembles the decision making process as done by human to a greater extent than different vector space models which only uses 'bag of words' concept.
Score:02

9. This results in "semantic loss" because semantic contextual senses of the sentences are discarded.
Score:01

Figure 6. Sentence Scoring

Single Research Paper Summary

Document ID:OSSIP01-19:01 Threshold:02

Research Goals]

In this paper, we present a new approach that incorporates semantic information from a document, in the form of Hierarchical Document Signature (HDS), to measure semantic similarity between sentences.
In this paper, we propose an application of hierarchical document signature (HDS), extension of fuzzy signature that takes into account semantic structure of sentences to measure sentences similarity.

[Research Methods]

Our HDS uses the natural hierarchy of the document and represents it in a modularized form of document level to sentence level, sentence to word level; aggregating similarity components at the lower levels and propagating them to the next higher level to produce the final similarity between sentences.
This poses a significant impact on many text mining applications performance where sentence-level judgment is involved.

[Research Continuation/Novel]

Due to variability of expressions of natural language, it is very essential to exploit the semantic properties of a document to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different.

[Research Similarity/Dissimilarity]

Each imparts different information to the context of a sentence.

[Research Outcome]

The evaluation of our HDS model has shown that it resembles the decision making process as done by human to a greater extent than different vector space models which only uses 'bag of words' concept. (HDS), to measure semantic similarity between sentences.

Figure 7. Single Paper Summary

Multiple Research Paper Summary

Document ID:OSSIP01-19:01,02 OSSIP01-19:02 Threshold:01

Document ID:OSSIP01-19:01

[Research Goals]

In this paper, we present a new approach that incorporates semantic information from a document, in the form of Hierarchical Document Signature (HDS), to measure semantic similarity between sentences.

[Research Methods]

Our HDS uses the natural hierarchy of the document and represents it in a modularized form of document level to sentence level, sentence to word level; aggregating similarity components at the lower levels and propagating them to the next higher level to produce the final similarity between sentences.

[Research Continuation/Novel]

Due to variability of expressions of natural language, it is very essential to exploit the semantic properties of a document to accurately identify semantically similar sentences since sentences conveying the same fact or concept may be composed lexically and syntactically different.

[Research Similarity/Dissimilarity]

Each imparts different information to the context of a sentence.

[Research Outcome]

The evaluation of our HDS model has shown that it resembles the decision making process as done by human to a greater extent than different vector space models which only uses 'bag of words' concept.

Document ID:OSSIP01-19:02

[Research Goals]

Based on the information people get from a sentence, *Objects-Specified Similarity*, *Objects-Property Similarity*, *Objects-Behavior Similarity* and *Overall Similarity* are defined to determine sentence similarities from four aspects.

[Research Methods]

Experiments show that the proposed method makes the sentence similarity comparison more exactly and give out a more reasonable result, which is similar to the people's comprehension to the meanings of the sentences.

[Research Continuation/Novel]

To deliver the sentence meaning more exactly, nowadays more and more applications require not only comparing the overall similarity between sentences but also the similarity between parts of these sentences.

[Research Outcome]

Experiments show that the proposed method makes the sentence similarity comparison more exactly and give out a more reasonable result, which is similar to the people's comprehension to the meanings of the sentences.

Figure 8. Multi-Paper Summary for 2 Input Articles.

V. SYSTEM EVALUATION

To determine how each article is relevant to readers' need, we used *F-measure* [10] performance matrix. The research article representing all five categories of RRN terms will be given a score of three; four categories a score of two; with a score of one. Other than above is assigned a score of zero. *F-measure* is calculated as,

$$F - measure = (2 * Precision * Recall) / (Precision + Recall) \quad (8)$$

We compared *OSS* results with human made summary for precision, recall d F-measure values as shown in Table II. Table III shows overall *F-measure* for all RRN categories.

TABLE II. Performance Per Category: F-measure (F), Precision (P), Recall(R).

	Research Goals			Research Methods			Research Similarity/ Dissimilarity			Research Continuation /Novel			Research Outcome		
	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R
AUTOMATED	52	45	65	26	34	20	61	57	66	86	84	88	45	40	50
HUMAN	63	72	56	52	50	55	79	79	79	93	94	92	71	68	75

TABLE III. F-Measure for Automatic & *OSS* Summary

RRN Categories	*F*-measure	
	Automatic Summary	Human Summary
Research Goal	62%	51%
Research Method	51%	28%
Research Contrast, Like	79%	60%
Research Continuation	92%	86%
Research Outcome	71%	45%

VI. CONCLUSION & FUTURE SCOPE

The new researcher or scholar as readers only searches for the most related published research articles of their interested areas for latest research developments in the same field. Since

the *Abstract* and *Introduction* sections of the research articles outlines the complete research, these two sections' summarized contents are sufficient to decide whether reader should read the complete paper to move ahead or not? In addition following are the advantages of *OSS*:

- The *OSS* evaluations show steady correlation with the human assessment outcome.
- Descending order sentence score comparison gives the most optimized results.
- Optimization enhances the effectiveness of *OSS* to make it function at its best and give its best advantage.
- Reduction significantly improves the conciseness of *OSS*.
- *OSS* introduces new information criteria by the use of dividing document into research categories thus providing new structure for summarization which improved output readability as compared with other systems.

Thus the *OSS* result informs that *OSS* is an effective and efficient strategy for providing short, condensed, accurate, explicit ,optimized and most related multiple research paper's summary, minimizing readers' efforts deciding whether to go ahead with the retrieved articles for further readings helping in his/her own work.

The *OSS* can further be modified for future improvements as a wide range of different sub-domains of the specific field can be covered. With only five RRN categories, it is not designed to model the full complexity of all research articles therefore RRN categories can be further subdivided for simplicity. All sections of the research papers can be covered to produce in depth summary but this need standardized formatting of topic, sub-topic headings. Citation text can be included to improve summary contents.

REFERENCES

[1]. Angrosh, M. A., Stephen Cranefield, and Nigel Stanger, "Context identification of sentences in related work sections using a conditional random field: towards intelligent digital libraries." JCDL'2010: Proceedings of the 10th annual JCDL, pages 293–302, ACM.
[2]. Alonso,Castellon, Climent (2002-2003), "Comparative study of Automated Text Summarization Systems".
[3]. Aone, Okurowski and Gorlinski , "Trainable Scalable Summarizer Using Robust NLP and Machine Learning", COLING, ACL,1998, pages62-66.
[4]. Cong Duy Vu Hoang and Min-Yen Kan, Towards Automated Related Work Summarization, 2010.
[5]. Goldstein, J, J. Carbonell ," The use of MMR diversity based re-ranking for reordering documents and producing summaries", In proceedings of SIGIR 1998, Melbourne, Australia .p335-336
[6]. Inderjeet Mani, "Summarization Evaluation: An Overview", Proceedings of the NTCIR Workshop 2, 2001.
[7]. Kan Min-Yen and McKeown, "Information Extraction and Summarization: Domain Independent through focus Types, Technical Report", Columbia University,NY, 1998.
[8]. King, Donald F.," Economic cost model of scientific scholarly journal publishing". ICSU Press Workshop, Keble College, University of Oxford, UK, 1998.
[9]. Kintsch, Walter, "The representation of meaning in memory. The Experimental psychology series. Lawrence Erlbaum Associates Publishers, 1974.
[10]. Kupiec, Julian, Jan O. Pedersen, and Francine Chen, "A trainable document summarizer", 18th Annual International Conference on Research and Development in IR (SIGIR-1995), pages 68–73.
[11]. Kripendorff , "Content Analysis: an introduction to its methodology", Sage commtext series; 5, Sage, Beverly Hills London, 1980.
[12]. Lapalme,Saggion ," Concept Identification and Presentation in the context of Technical Text Summarization", 2001.
[13]. Lapalme, Saggion , "Evaluation of content and text quality in the context of Technical Text Summarization", RIAO' 2000, Paris, France.
[14]. Lin, Chin- Yew , "Training a selection function for Extraction", In ACM, CIKM, pages 55-62, 1999.
[15]. Mahajan Sunita, Patil Sunita, "Cluster Based Sentence Extraction Approach for Summarizing Published Research Paper Abstracts", IC4E, Mumbai, India, 2011.
[16]. Mani Inderjeet and Maybury, M., editors, "Advances in Automatic Text Summarization". MIT Press, 1999.
[17]. Mani Inderjeet, Barbara Gates, and Eric Bloedorn," Improving summaries by revising them", ACL-99, pages 558–565, 1999.
[18]. Baxendale, P,"Machine-made index for technical literature - an experiment. IBM Journal of Research Development", 2(4) :{354-361}.
[19]. Buckley, Christopher," Implementation of the SMART information retrieval system". Technical Report 85-686, Cornell University, 1985.