International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:2, 2015

Automatic Enhanced Update Summary Generation System for News Documents

S. V. Kogilavani, C. S. Kanimozhiselvi, S. Malliga

***Abstract*—**Fast changing knowledge systems on the Internet can be accessed more efficiently with the help of automatic document summarization and updating techniques. The aim of multi-document update summary generation is to construct a summary unfolding the mainstream of data from a collection of documents based on the hypothesis that the user has already read a set of previous documents. In order to provide a lot of semantic information from the documents, deeper linguistic or semantic analysis of the source documents were used instead of relying only on document word frequencies to select important concepts. In order to produce a responsive summary, meaning oriented structural analysis is needed. To address this issue, the proposed system presents a document summarization approach based on sentence annotation with aspects, prepositions and named entities. Semantic element extraction strategy is used to select important concepts from documents which are used to generate enhanced semantic summary.

***Keywords*—**Aspects, named entities, prepositions, update summary.

I. Introduction

IN recent times, online web content data is made available at an increasing speed and people prefer to develop a crisp overview from a large number of articles in quick time. So, document summarization aims at generating concise, comprehensible and semantically meaningful summaries. Producing updated information could be valuable for people to use the latest information by eliminating the surplus data.

The proposed system aims to produce semantic summary by using a list of important aspects such as what, when, how etc. These lists of aspects define what counts as important information but the summary also includes other facts which are considered equally important. The summary should contain the aspect oriented information for all aspects. The summary generated is guided by predefined aspects that are employed to enhance the quality and readability of the resulting summary.

The guided summarization task is to create a summary of a set of ten newswire articles for a given topic, where the topic falls into a predefined category. Human summarizers are given a list of important aspects for each category, and a summary must cover all these aspects. The summaries may also contain information relevant to the topic. It guides the systems to do so by explicitly assigning topics to categories and advocating the systems to retrieve content relevant to the aspects of each category. Some of the categories are accidents and natural disasters, attacks, health and safety, investigation and trails. Each category includes aspects such as what, when, where, why, who affected, damages, countermeasures, perpetrators, threats, plead, sentence and charges etc.

The resultant summary should cover all the relevant aspects if such information can be found in the source documents. The summary should also include other relevant information, if it's crucial to the topic. A possible starting strategy for summary generation is sentence extraction. An extraction system tries to select the sentences with the most important information from the documents. These sentences can simply be presented in full to a user as an indicative summary.

Ideally, the documents would be thoroughly analyzed using linguistic and world knowledge to determine which sentences are appropriate for the extract. Many existing systems extract sentences on the basis of a limited set of mundane features. The proposed Enhanced Update Summary Generation (EUSG) approach enhances the existing feature based extraction strategy by including semantic feature.

II. Literature Review

Due to the rapid evolution of information quantum on the Internet, update summarization has received much attention in recent years. It seeks to summarize an evolutionary document collection at current time on the supposition that users have read some related previous documents.

According to [1], multi-document summarizer, MEAD generates summaries using cluster centroids which were produced by a topic detection and tracking system. A key feature of MEAD was its use of cluster centroids, which consist of words that are central not only to one article in a cluster, but to all the articles. It used information from the centroids of the clusters to select sentences that are most likely to be relevant to the cluster topic. But with the limited data set, this method could not claim any statistical significance.

The system proposed in [2] used simple techniques derived from Latent Semantic Analysis (LSA) to provide a simple and robust way of generating extractive summaries for update summarization task.

Reference [3] developed a summarizer which is based on Iterative Residual Rescaling (IRR) that created the latent semantic space for a set of documents under consideration. IRR generalized Singular Value Decomposition (SVD) and enabled the control of the influence of major and minor topics in the latent space.

The approach specified in [4] described a method for multi-document update summarization. The best summary was

S.V Kogilavani is with the Kongu Engineering College, Erode, TN 638052 India (phone: 09486153223; e-mail: vani_sowbar@yahoo.co.in).

C.S Kanimozhiselvi and S. Malliga are with the Kongu Engineering College, Erode, TN 638052 India (e-mail: kanimozhi@kongu.ac.in, mallisenthil@kongu.ac.in).

defined to be the one which had the minimum information distance to the entire document set. The best update summary has the minimum conditional information distance to a document cluster given the fact that a prior document cluster had already been read.

Reference [5] proposed an update summarization framework based on topic correlation analysis. The topics were first extracted from the two document sets provided in the task of update summarization by means of Latent Dirichlet Allocation (LDA) topic model. Then, the correlation between the new topics and the old topics were identified, based on which four categories of topic evolution patterns were defined to capture the topic shift between the two document collections. A new sentence ranking algorithm, CorrRank was developed, which fully incorporates the topic evolution in the process of sentence ranking and sentence selection in update summarization. Reference [6] proposed a new method based on shallow parsing with rules. The rules were generated according to the syntactic features of English texts, such as the tense of verbs, the usages of modal verbs and so on. The latest novel information from English news texts was extracted correctly, to meet the needs of users for accessing of updated information of the developing events quickly and effectively.

Reference [7] proposed a multi-document summarization, where aspects could be taken as specified queries in summarization. They also proposed a novel ranking algorithm, Decayed DivRank for guided summarization tasks. This could address relevance, importance, diversity, and novelty simultaneously through a decayed vertex-reinforced random walk process in sentence ranking.

The concept of hierarchical topic for multi-document automatic summarization task was proposed in [8], which used multi-layer topic tree structure to represent the text set. Each node in the topic tree represented a specific topic and contained multiple similar sentences in the text set. The hierarchical topic structure could describe accurately the similarity between sentences at different levels of granularity. Therefore it could reflect the real content of the text set than a single layer topic set.

A NEws Symbolic Summarizer (NESS) system [9] was developed to rely on the syntactical parser to extract linguistic knowledge from source documents. It selects sentences based on linguistic metrics. It also measures the similarity between candidate sentences and the previous articles already read by the user. It utilizes Term Frequency-Inverse Document Frequency (TF-IDF) to measure the relevance of the sentence to the topic. This NESS system is considered as Baseline system for evaluation.

III. Proposed System Overview

A collection of topic related two sets of documents are fed as input. The output is a concise set of two summaries that contain reduced information. The main aim is to simulate a user who is interested in learning about the latest developments on a specific topic and who wishes to read a brief summary of the latest news. The proposed system design is represented in Fig. 1.

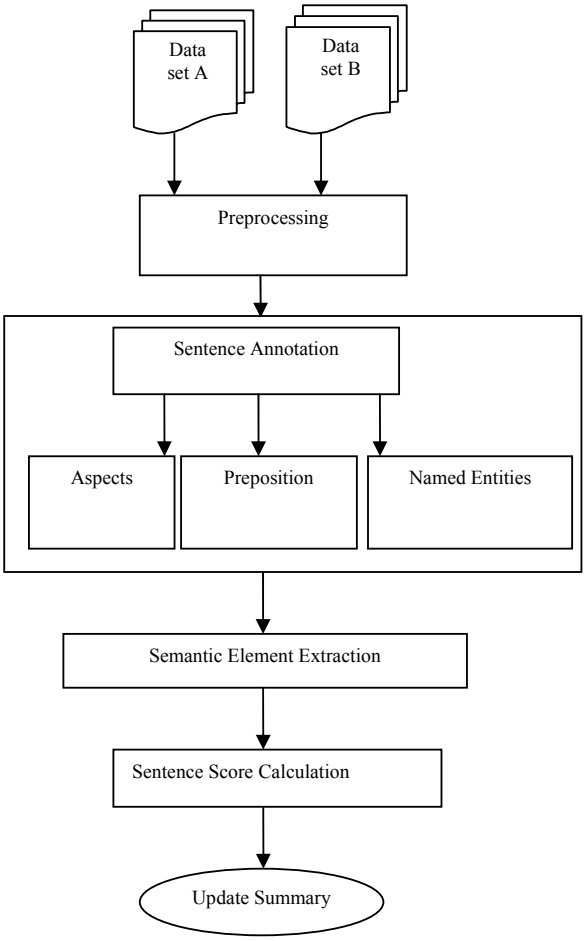

Fig. 1 Proposed System Overview

A. Enhanced Summary Generation Steps

Step 1. Initially the articles in the dataset are split into sentences and those sentences are annotated with predefined aspects, prepositions and named entities.

Step 2. Sentence representation is enhanced by extracting concepts from Wikipedia, which is referred to as sentence wikification process.

Step 3. Individual sentences are mapped into concepts and individual word score is calculated based on TSCF-ISF measure.

Step 4. Then for each sentence, score is calculated based on basic and additional features for dataset A articles and based on basic, additional as well as update features for dataset B articles.

Step 5. Highest ranking sentences are selected and ordered in a way in which the sentences are included in the original documents and finally initial summary is generated.

Step 6. Update summary is generated after removing redundancy.

B. Sentence Annotation

In order to select salient sentences, all the sentences are annotated with predefined aspects, prepositions and named

entities. The articles from the dataset are split into sentences and annotated with appropriate template tags. These annotations include both objective (when, where, who) and subjective (how, why, countermeasures) tags. As any standard Named Entity Recognition (NER) can only tag objective tags, the proposed EUSG system would manually annotate all the articles with all possible tags. A sentence is tagged with multiple tags if it has more than one answer to the template.

- *Sentence Annotation with Aspects*

Consider the sentence taken from the document D08021D:NYT_ENG_20050707 related to Attacks category. Table I denotes sample sentences and sentences annotated with aspects.

TABLE I
ASPECTS BASED ANNOTATION

Sample Sentence	Sentence Annotated with Aspects
The next explosion occurred at 8:56 a.m. near King's Cross Station, where the death toll was 21, the police said.	The next explosion occurred at <when>8:56 a.m.<when> near <where>King's Cross Station</where>, where the <who affected>death toll was 21</who affected>, the police said.
Twenty-one minutes later, at 9:17 a.m., a third blast ripped through a train coming in Edgware Road underground station, killing seven.	Twenty-one minutes later, at <when>9:17 a.m.</when>, a third blast ripped through a train coming in <where>Edgware Road</where> underground station, <who affected>killing seven</who affected>.

- *Sentence Annotation with Prepositions*

In English grammar, a preposition is a part of speech that links nouns and pronouns to other phrases in a sentence. A preposition generally represents the temporal, spatial or logical relationship of its object to the rest of the sentence. It is very interesting to observe how prepositions implicitly capture the key elements in a sentence. The list of prepositions used for calculating sentence importance are limited to simple single word prepositions like "in", "on", "of", "at", "for", "from", "to", "by", "with" etc. Annotations of the sentences with prepositions are given in Table II.

TABLE II
PREPOSITIONS BASED ANNOTATION

Sample Sentence	Sentence Annotated with Prepositions
The next explosion occurred at 8:56 a.m. near King's Cross Station, where the death toll was 21, the police said.	The <preposition>next</preposition> explosion occurred <preposition>at</preposition> 8:56 a.m. <preposition>near</preposition> King's Cross Station, where the death toll was 21, the police said.
Twenty-one minutes later, at 9:17 a.m., a third blast ripped through a train coming in Edgware Road underground station, killing seven.	Twenty-one minutes later, <preposition>at</preposition> 9:17 a.m., a third blast ripped <preposition>through</preposition> a train coming <preposition>in</preposition> Edgware Road underground station, killing seven.

- *Sentence Annotation with Named Entities*

Prior observations in the given data led to the understanding that more the types of named entities a sentence contains, the stronger is the likelihood of the sentence's capabilities in answering a set of questions like What happened? Who was involved? And where did this happen?. Named entities refer to the objects for which proper nouns are used in a sentence. Seven basic named entities are identified such as person, location, date, time, organization, money and percentage. Stanford Named Entity Recognition discussed in [10] employs person, location, organization entities. Others are extracted by applying patterns. Annotations of the sentences with named entities are given in Table III.

TABLE III
NAMED ENTITIES BASED ANNOTATION

Sample Sentence	Sentence Annotated with Named Entities
The next explosion occurred at 8:56 a.m. near King's Cross Station, where the death toll was 21, the police said.	The next explosion occurred at <time>8:56 a.m.</time> near <location>King's Cross Station</location>, where the death toll was 21, the <person name>police</person name> said.
Twenty-one minutes later, at 9:17 a.m., a third blast ripped through a train coming in Edgware Road underground station, killing seven.	Twenty-one minutes later, at <time>9:17 a.m.</time>, a third blast ripped through a train coming in <location>Edgware Road</location> underground station, killing seven.

C. Semantic Element Extraction

The words are conventionally considered to be units of the text to calculate importance. Simple word counts, frequencies and synonym based word frequencies in the document collection have proved to work well in the context of summarization in [11] The proposed EUSG system uses semantic concepts in computing sentence importance. Wikipedia is a vast, interlinked network of articles providing multilingual database of concepts. It is a web based, free content encyclopedia, comprehensive and well organized knowledge repository defined in [12]. Links are there in Wikipedia's articles which are used to direct the user to recognize related pages. Wikipedia Miner is a freely available toolkit for navigating and making use of the content of Wikipedia. The proposed EUSG system seeks to create a concept database from Wikipedia concepts by selecting the concepts that appear explicitly in sentences. Each word in each sentence is compared with the concept database.

Let Con = $\{cp_1, cp_2, \ldots, cp_n\}$ be the set of concepts in the concept database. To improve accuracy and to calculate the weight of each word, the proposed EUSG system adopts Term Synonym Concept Frequency (TSCF). Every word's TSCF is calculated by performing synset extraction, concept database construction and term frequency calculation. Word weight is calculated as defined in (1):

$$\text{Word} - \text{Weight}(w_i) = \frac{\text{TSCF}(w_i) \times \text{ISF}(w_i)}{K}. \quad (1)$$

Here i = 1, 2, ..., M. TSCF of every word is obtained by using (2):

$$\text{TSCF}(w_i) = \sum_{i=1} \alpha \times \text{TF}(w_i) + \beta \quad (2)$$

where $w_i \in \{w \cup \text{syn}(w)\}$. In TSCF calculation, to include word synonym into account, the TF denotes Term Frequency of each word and it's synonym is multiplied by α where α = 1

for the word and $\alpha = 0.5$ for the synonym of word and $\beta = 1$ if the word itself is a concept in the concept database.

IV. Update Summary Generation

Update summary is generated by using Novel Sentence Similarity Measure. This new feature selects novel sentences that have not been included in the initial summary. All sentences in the initial summaries are considered as candidate sentences. New sentences that have least similarity with these candidate sentences are chosen as sentences in the update summary. The similarity between candidate sentences and sentences in dataset B is calculated as in (3):

$$\text{NSSM}(S_i) = \text{Sim}(S_i, S_j) = \frac{\sum w_i}{\sum w_j} \quad (3)$$

where, $w_i \in S_i \cap S_j$, $w_j \in S_{min}$. The numerator is the sum weight of the words that occur both in sentence S_i and S_j. The denominator is the sum weight of the words that occurs in the shorter sentence S_{min} in $\{ S_i, S_j \}$.

The benefit is that if a sentence contains all the words of another sentence, i.e. if one sentence is totally a part of another, then their similarity is 1. Now the sentence score is calculated for all sentences using update feature.

V. Experimental Result Analysis

A. Evaluation Domain

Summarization methods will be evaluated on the TAC 2008 dataset which is useful for summarization task. The documents for summarization are taken from the TAC 2008 AQUAINT-2 collection of newswire articles. The AQUAINT-2 collection comprises news articles spanning the time period of October 2004-March 2006. Articles are in English with 48 topics and each topic consists of 20 documents and is divided into two sets of 10 documents each, such that dataset B followed dataset A in the temporal order.

B. Recall / Precision

The comparison between summaries can be carried out by humans, but it can often be computed automatically. A variety of different measures can be used for evaluation. Relevancy is often measured using IR metrics such as Precision and Recall as in [13]. In pattern recognition and information retrieval, Precision is the fraction of retrieved instances that are relevant, while Recall is the fraction of relevant instances that are retrieved. Both Precision and Recall are therefore based on an understanding and measure of relevance.

Recall is the ability of the search to find all the relevant items in the corpus. It is defined as the fraction of the sentences that are relevant to the topic that are successfully retrieved.

$$\text{Recall} = \frac{(\text{Relevant} \cap \text{Retrieved})\text{Sentences}}{\text{RelevantSentences}} \quad (4)$$

In (4), relevant sentences are sentences that are identified in the human generated summary and retrieved sentences are sentences that are retrieved by the system. Precision is the ability to retrieve top-ranked sentences that are mostly relevant. It is defined as the fraction of retrieved sentences that are relevant to the search as in (5). A higher Precision implies that most relevant sentences are selected but includes lots of junk. The higher Recall value indicates that the system returns relevant sentences but misses many useful ones too.

$$\text{Precision} = \frac{(\text{Relevant} \cap \text{Retrieved})\text{Sentences}}{\text{RetrievedSentences}} \quad (5)$$

The trade-off between Precision and Recall is that a high value of Precision returns relevant sentences whereas a high value of Recall returns most relevant sentences. Another factor that must be considered is the total number of sentences. It is often possible to achieve high Precision and high Recall rates in a small corpus, but as the corpus size increases, these rates drop considerably. To maintain standard Recall level, a Precision value is interpolated for each standard Recall level. The intersection of the Precision and Recall point is a critical optimization data point, and IR systems attempt to move this data point closer to the top right corner in graph representation.

C. ROUGE-1 Score

Human judgment often has wide variance on what is considered a good summary, which means that making the evaluation process automatic is particularly difficult. Manual evaluation can be used, but this is both time and labor intensive as it requires humans to read not only the summaries but also the source documents. One metric used is the Recall-Oriented Understudy for Gisting Evaluation (ROUGE). It essentially calculates n-gram overlaps between automatically generated summaries and previously written human summaries as in [14]. A high level of overlap should indicate a high level of shared concepts between the two summaries.

Automated machine summaries can be compared with human summaries using the ROUGE summarization evaluation tool. ROUGE scores range from 0 to 1 and it reflects the similarity, with a higher score reflecting more similarity between two summaries. The ROUGE-1 score is based on the overlap of unigrams between automatically generated summaries and human generated summaries and it solely reflects the overlap in vocabulary between two summaries.

$$\text{ROUGE}-1\text{Score} = \frac{C}{T} \quad (6)$$

In (6), C is the count of number of unigrams that occurs in machine and human summary and T is total number.

Fig. 2 shows the word score calculated by standard TF-IDF and proposed TSCF-ISF measure. The result indicates that improved accuracy is obtained by the TSCF-ISF measure.

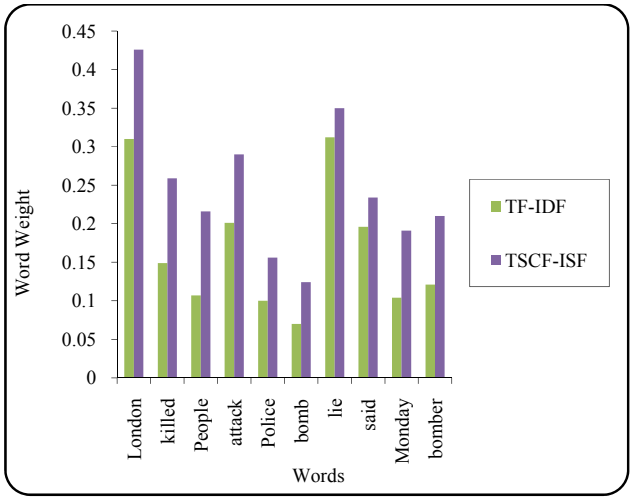

Fig. 2 Word Weight Measure Analysis

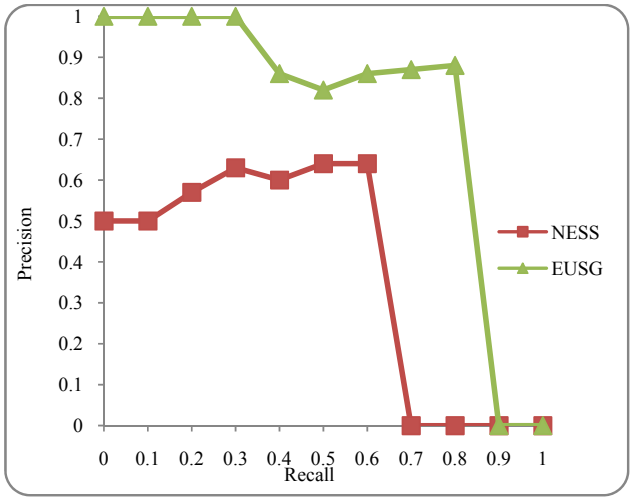

Fig. 3 Recall-Precision Graph of Update Summary at 20% CR

TABLE IV
EUSG ACTUAL RECALL AND PRECISION AT 20% CR

Relevant Sentences (Yes/No)	Precision	Recall
Yes	1.00	0.07
Yes	1.00	0.13
Yes	1.00	0.20
Yes	1.00	0.27
Yes	1.00	0.33
No	0.83	0.33
Yes	0.86	0.40
No	0.75	0.40
Yes	0.78	0.47
Yes	0.80	0.53
Yes	0.82	0.60
Yes	0.86	0.67
Yes	0.87	0.73

Table IV shows actual Recall and Precision calculated by proposed EUSG system for update summary at 20% CR.

Fig. 3 shows Recall - Precision graph at 20% CR for the update summary. This figure shows that at 20% CR, high Precision is obtained by the proposed EUSG system.

Table V shows interpolated Recall and Precision calculated by proposed EUSG system for update summary at 20% CR.

TABLE V
EUSG INTERPOLATED RECALL AND PRECISION AT 20% CR

Precision	Recall
0.0	1.00
0.1	1.00
0.2	1.00
0.3	1.00
0.4	0.86
0.5	0.82
0.6	0.86
0.7	0.87
0.8	0.88
0.9	0.00
1.0	0.00

Fig. 4 compares the ROUGE-1 Score of different summaries. The result shows that the overlap between Initial Summary (IS) and Update Summary (US) is low in the proposed EUSG system. Also IS and US of the proposed EUSG system highly correlate with human summary when compared to the existing NESS system.

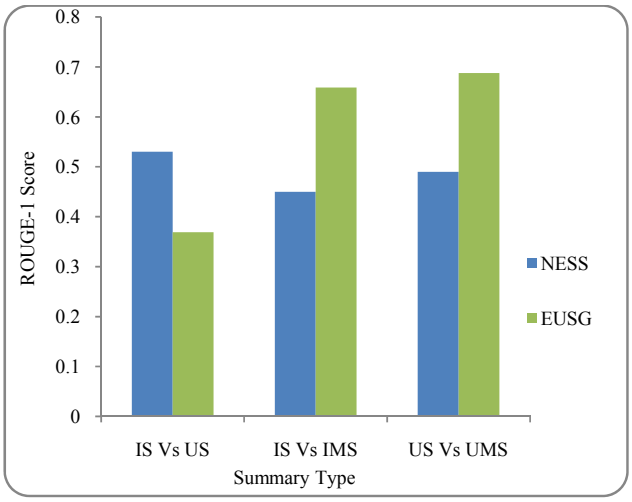

Fig. 4 ROUGE-1 Score Performance Measure

VI. CONCLUSION AND FUTURE WORK

The proposed system generates initial and update summary from multiple documents based on annotating the sentences and relevant sentences are selected by utilizing Wikipedia which is used to get concepts and by applying different combinations of features. The relevancy is improved by adopting TSCF-ISF measure. The update summary generated by applying the proposed NSSM measure is compared with manual summary as well as with its initial summary and the result shows that the proposed system summary is proficient compared to existing NESS system.

For multi-document summarization systems, it is important to determine a coherent arrangement of the textual segments. A summary with improperly ordered sentences confuses the reader and degrades the quality or reliability of the summary itself. Further research may focus on this sentence ordering strategy. Sentence clustering was recently explored in research literature in order to provide more informative summaries. Existing cluster based ranking approaches applied clustering and ranking in isolation. Significant future attention may be paid on this aspect and an approach is needed that tightly integrates ranking and clustering by mutually and simultaneously updating each other so that the performance of both could be improved.

REFERENCES

[1] Dragomir Radev, R., Hongyan Jing, Malgorzata Stys and Daniel Tam, "Centroid-based summarization of multiple documents", International Journal of Information Processing and Management, Vol. 40, pp. 919-938, 2004.
[2] Kirill Kireyev, "Using latent semantic analysis for extractive summarization", In Proceedings of Text Analysis Conference, 2008.
[3] Josef Steinberger and Karel Jezek, "Update Summarization Based on Novel Topic Distribution", In Proceedings of the 9th ACM Symposium on Document Engineering, pp. 205-213, 2009.
[4] Chong Long, Min-Lie Huang and Xiao-Yan Zhu, "A New Approach for Multi-Document Update Summarization", Journal of Computer Science and Technology, Vol. 25, No. 4, pp. 739-749, 2010.
[5] Lei Huang and Yanxiang He, "CorrRank: Update Summarization Based on Topic Correlation Analysis", Lecture Notes in Computer Science, Vol. 6216, pp. 641-648, 2010.
[6] Min Peng, Xiaoxiao Ma, Ye Tian, Hua Long, Quanchen Lin and Xiaojun Xia, "The Web Information Extraction for Update Summarization Based on Shallow Parsing", In Proceedings of International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, pp. 109-114, 2011.
[7] Pan Du, Jipeng Yuan, Xianghui Lin, Jin Zhang, Jiafeng Guo and Xueqi Cheng, "Decayed DivRank for Guided Summarization", In Proceedings of Text Analysis Conference, 2011.
[8] Chandra, M., Gupta, V. and Paul, S.K. "A Statistical Approach for Automatic Text Summarization by Extraction", In Proceedings of International Conference on Communication Systems and Network Technologies, pp. 268-271, 2011.
[9] Pierre-Etienne Genest, Guy Lapalme, Luka Nerima and Eric Wehrli, "A symbolic Summarizer for the Update Task of TAC 2008", In Proceedings of Text Analysis Conference, 2008.
[10] Jenny Rose Finkel, Trond Grenager and Christopher Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 363-370, 2005.
[11] Vivi Nastase, David Milne and Katja Filippova, "Summarizing with Encyclopedic Knowledge", In Proceedings of Text Analysis Conference, 2009.
[12] Yajie, M. and Li, C. "WikiSummarizer - A Wikipedia-based summarization system", In Proceedings of the Text Analaysis Conference, 2010.
[13] Christopher Manning, D., Prabhakar Raghavan and Hinrich Schutze, "Introduction to Information Retrieval", Cambridge University Press, Cambridge, 2008.
[14] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In Proceedings of the ACL-04 Workshop, pp. 74-81, 2004.

Kogilavani is born in Coimbatore, TN, in the year 1978. She completed her B.E (Computer Science and Engineering) in the year 1999 from Madras University, Chennai and obtained M.E (Computer Science and Engineering) degree in the year 2007 from Anna University, Chennai. She got her Ph.D from Anna University, Chennai in the year 2013. Her research area is information retrieval, summarization and opinion mining.

She is associated with the Department of Computer Science and Engineering as Senior Assistant Professor at Kongu Engineering College, Tamil Nadu, India .She has presented 15 papers in national and international conferences and published 10 papers in national and international journals. She conducted many courses for the benefits of students . She conducted various workshops and seminars. She has also guided many UG and PG projects.

Dr. Kogilavani Shanmugavadivel is life member of Computer Society of India. She got academic excellence award from Kongu Engineering College in the year 2005.