Automatic Multi Document Summarization Approaches

[1]Yogan Jaya Kumar and [2]Naomie Salim
[1]Faculty of Information and Communication Technology,
University Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia
[2]Faculty of Computer Science and Information Systems,
University Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

Abstract: Problem statement: Text summarization can be of different nature ranging from indicative summary that identifies the topics of the document to informative summary which is meant to represent the concise description of the original document, providing an idea of what the whole content of document is all about. **Approach:** Single document summary seems to capture both the information well but it has not been the case for multi document summary where the overall comprehensive quality in presenting informative summary often lacks. It is found that most of the existing methods tend to focus on sentence scoring and less consideration is given to the contextual information content in multiple documents. **Results:** In this study, some survey on multi document summarization approaches has been presented. We will direct our focus notably on four well known approaches to multi document summarization namely the feature based method, cluster based method, graph based method and knowledge based method. The general ideas behind these methods have been described. **Conclusion:** Besides the general idea and concept, we discuss the benefits and limitations concerning these methods. With the aim of enhancing multi document summarization, specifically news documents, a novel type of approach is outlined to be developed in the future, taking into account the generic components of a news story in order to generate a better summary.

Key words: Multi document summarization, extractive summarization, news documents, summarization approaches, ontology learning, generic components, novel text summarization approaches

INTRODUCTION

The need of automatic text summarization has recently increased due to the proliferation of information on the Internet. With the availability and speed of internet, information search from online documents has been eased down to user's finger tips. However, it is not easy for users to manually summarize those large online documents. For example, when a user searches for information about earthquake which occurred in Sendai, Japan, the user will probably receive enormous articles related to that event. The user would definitely opt for a system that could summarize those articles. The goal of automatic text summarization is condensing the source text into a shorter version preserving its information content and overall meaning.

The objective and approach of summarization of documents explain the kind of summary that is generated. For example, it could be indicative of what a particular subject is about (closely related to a user query), or can be informative about what the whole content of document is all about. Besides that, approach towards text summarization can be either extractive or abstractive (Radev *et al.*, 2002). In extractive type summarization, important sentences are identified and directly extracted from the original document, i.e. the final summary consists of original sentences. On the other hand, in abstractive type summarization (Ganesan *et al.*, 2010) the sentences which are selected from the original document are further processed to restructure them before concatenating them into final summary. This process usually involves deep natural language analysis and sentence compression.

By understanding the type of summary i.e., indicative, informative, extractive and abstractive, we can then apply them to either single document or multi document. This study focuses mainly on informative and extractive type multi document text summarization. The distinct characteristics that make multi document summarization rather different from

Corresponding Author: Yogan Jaya Kumar, Faculty of Information and Communication Technology,
University Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia

single document summarization is that multi document summarization problem involves multiple sources of information that overlap and supplement each other, being contradictory at occasions. So the key tasks are not only identifying and coping with redundancy across documents, but also ensuring that the final summary is both coherent and complete.

The contributions of this study can be summarized as follows: We discuss the four notable approaches of multi document summarization and present it with related research from literatures. The benefits and limitations concerning these approaches are also discussed. At the end of this study, a novel type of approach is outlined to be developed in the future, for news documents summarization. We aim to incorporate the generic components of a news document in order to generate a better summary.

The rest of the study is organized as follows: First we present the survey on four multi document summarization approaches namely the feature based method, cluster based method, graph based method and knowledge based method. Then we outline the proposed multi document summarization method; i.e., the component based method. Finally we end with conclusion.

Multi document summarization approaches: A number of research study have addressed multi document summarization in academia (Erkan and Radev, 2004a, Wan and Yang, 2008, Haribagiu and Lacatusu, 2010) and illustrated different types of approaches and available systems for multi document summarization. In this study we direct our focus notably on four well known approaches to multi document summarization. Our discussion will be based on the following pattern: For each method, we will first discuss its main idea. Following that, we will look at some research study from related literatures. Finally the benefits and limitations concerning each method are commented.

Feature based method: Extractive type summarization involves identifying the most relevant sentences from the text and put them together to create a concise summary. In the process of identifying important sentences, features influencing the relevance of sentences are determined. Here we list some of the common features that have been considered for sentence selection.

Word frequency: The idea of using word frequency is that important words appear many times in the document. The most common measure widely used to calculate the word frequency is tf and idf.

Title/headline word: Occurrence of words from the document title in sentence indicates that the sentence is highly relevant to the document.

Sentence location: Important information in a document is often covered by writers at the beginning of the article. Thus the beginning sentences are assumed to contain the most important content.

Sentence length: Very short sentences are usually not included in summary as they convey less information. Very long sentences are also not suitable to represent a summary.

Cue word: There are certain words in a sentence which indicate that the sentence is carrying an important message in the document (e.g., "significantly", "in conclusion").

Proper noun: Sentences containing proper noun representing a unique entity suchlike name of a person, organization or place are considered important to the document.

Let us show a simple example for scoring a sentence S. Consider that we select three features namely title word, sentence length and sentence location. Calculation of the score for each of these features as in (Suanmali *et al.*, 2009) is shown below:

$$\text{Title word}, T = \frac{\text{No. title word in S}}{\text{No. word in title}}$$

$$\text{Sentence length}, L = \frac{\text{No. word in S}}{\text{No. word in longest sentence}}$$

$$\text{Sentence position}, P = 5/5 \text{ for 1st, } 4/5 \text{ for 2nd,}$$
$$3/5 \text{ for 3rd, } 2/5 \text{ for 4th,}$$
$$1/5 \text{ for 5th, } 0/5 \text{ for other sentences}$$

After computing each of the feature score, the total scoring of a sentence is given by Eq. 1:

$$\text{Score\_S}_i = w_1 T_i + w_2 L_i + w_3 P_i \qquad (1)$$

where, Score_S_i is the total score of sentence S_i. The terms T_i, L_i and P_i are the feature scores of sentence S_i based on the title words it contains, sentence length and sentence position respectively. w_1, w_2 and w_3 are the weights for the linear combination of the three features. Figure 1 depicts the generalized architecture of a feature based summarizer.
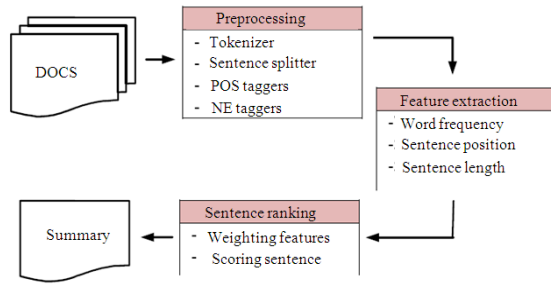
Fig. 1: Feature based summarizer generalized architecture

Table 1: Weight learning methods

Authors	Weight learning method
(Osborne, 2002)	Conjugate gradient decent search method
(Fattah and Ren, 2009)	Mathematical Regression (MR) model
(Binwahlan *et al.*, 2009)	Particle Swarm Optimization (PSO)
(Dehkordi *et al.*, 2009),	Genetic Algorithm (GA) model
(Bossard and Rodrigues, 2011)	
and (Suanmali *et al.*, 2011)	

We can see that text features score are combined for sentence scoring, as shown in Eq. 1. But not all text features are treated with same level of importance as some of the features have more importance or weight and some have less. Thus emphasis should be given on dealing with the text features based on their importance. This issue can be overcome by using weight learning method. Many researchers have been using various weight learning methods in their study. Binwahlan *et al*. (2009) introduced a novel text summarization model based on swarm intelligence technique known as Particle Swarm Optimization (PSO). The feature scores were adjusted using the weights resulting from the training using the PSO. The training set consist pairs of document and human reference summary. They implemented the sentence scoring using the following Eq. 2:

$$Score(S) = \sum_{i=1}^{5} w_i \times score\_f_i(s) \qquad (2)$$

where, Score (S) is the score of the sentence S, w_1 is the weight of the feature i produced by PSO, i = 1-5 showing that 5 text features where used and Score_f_i(s) is the score of the feature i. The use of PSO for optimization have also been proven to be robust in other domains as well (Nacy *et al*., 2009, Balaji and Kamaraj, 2011). Another weight learning approach was described by Bossard and Rodrigues (2011) who approximated the best weight combination using a genetic algorithm for their multi document summarizer. By using the genetic algorithm, a suitable combination of feature weights can be found. Table 1 lists some of the weight learning methods applied to text summarization.

Besides optimizing feature weights, the impact of combining different features has been investigated by Hariharan (2010) for multi document summarization. In his study, the author showed that term frequency weight combined with position and node weight feature yields significantly better results.

The feature based method has been widely used by researchers because of its simple and direct approach to sentence selection for summarization. It is found that the combination of cue words, title words and location are what relied upon as primary features (Gupta and Lehal, 2010). An important issue to be noted in the context of multi document summary is that important or relevant information is usually spread across documents and feature based methods often fail to handle this problem. Besides that, the feature based method is knowledge poor in term of capturing contextual information contents that exist in the sentences and multiple documents. These limitations are due to the sentence scoring process which depends solely on flat feature representation of a sentence while omitting cross-document concepts and varying context in different documents.

Cluster based method: The idea of clustering is to group similar objects into their classes. As far as multi documents are concerned, these objects refer to sentences and the classes represent the cluster that a sentence belongs to. By looking at the nature of documents that address different subjects or topics in the documents, some researchers try to incorporate the idea of clustering into their study. Using the concept of similarity, sentences which are highly similar to each other are grouped into one cluster, thus generating a number of clusters. The most common technique to measure similarity between a pair of sentences is the cosine similarity measure where sentences are represented as a weighted vector of tf-idf. Once sentences are clustered, sentence selection is performed by selecting sentence from each cluster. Sentence selection is then based on the closeness of the sentences to the top ranking tf-idf in that cluster. Those selected sentences are then put together to form the final summary Fig. 2.
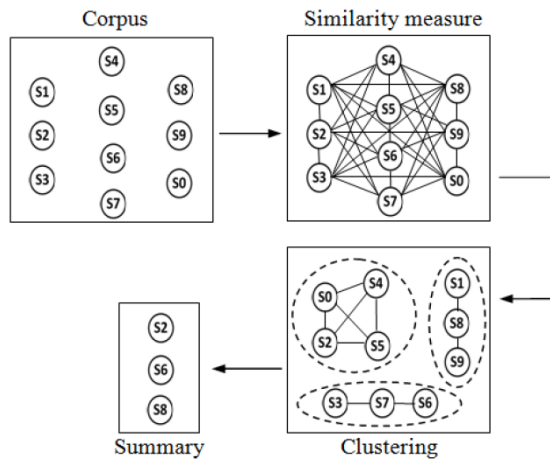
Fig. 2: Cluster based summarization

Typically, clustering algorithms can be categorized as agglomerative or partitional (Jain *et al.*, 1999). In agglomerative clustering (also known as "bottom-up" approach), each sentence is initially considered a separate cluster by its own. This individual clusters are then merged into successively larger clusters. This iterative process ends when some stopping criterion is reached.

Whereas in partitional clustering approach, initially all sentences are grouped into one big cluster. Iteratively smaller clusters are generated by dividing the largest cluster into several sub-clusters. Each sub-cluster generated will then contain sentences with higher similarity. A well known partitional clustering algorithm is the K-Means algorithm.

Radev *et al.* (2004) pioneered the use of cluster centroids for their multi-document summarizer, MEAD. Centroids are the top ranking tf-idf that represents the cluster. These cluster centroids are then used to identify the sentences in each cluster that are most similar to the centroid. Thus, the summarizer generates sentence which are most relevant to each cluster.

Taking the benefit of clustering approach, efforts have been put into making the overall process of summarizing multi document effective. One that is worth to be mentioned here is determining the optimal number of clusters, where Xia *et al.* (2011) adopted the co-clustering theory to find optimal clusters. They determine the weights of sentences and terms based on the sentence-term co-occurrence matrix. Sentence-term matrix is designed to represent diversity and redundancy within multiple articles.

Finally the top-weighted sentence in every cluster is picked out to form the summary, until a user-preferred summary length is met.

Some researchers employ clustering-based hybrid strategy (Yu *et al.*, 2006) to combine local and global search for sentence selection. This approach does not depend only on similarity to cluster for sentence selection but also considers the overall document content similarity. Focus has also been given on strengthening clusters diversity. To achieve this, Aliguliyev (2010) used PSO algorithm by adding a mutation operation adopted from genetic algorithms to optimize intra-cluster similarity and inter-cluster dissimilarity.

Cluster based methods has been successful in its task to represent diversity and reduce redundancy within multiple articles. Although these can be considered the advantage of using clustering methods, as far as multi document is concerned, a summary cannot be meaningful enough if the relevance of a sentence is judged merely based on the clusters. This is because in clustering based method, eventually sentences are ranked according to the similarity with cluster centroid which simply represents frequent occurring terms. Thus, this method is also considered to be knowledge poor in term of its inability to capture contextual information contents that exist in the sentences.

Graph based method: The fundamental theory of graph representation is the connection or linking between objects. These connections exist based on their underlying relation. In the case of text documents, the underlying relation is usually the similarity between objects-in this case, sentences.

Generally, a graph can be denoted in the form of G = (V, E), where V represents the graph's vertex or node and E is the edge between each vertex. In the context of text documents, vertex represents sentence and edge is the weight between two sentences. Using this approach, documents can therefore be represented as a graph where each sentence becomes the vertex and the weight between each vertex corresponds to the similarity between the two sentences. As in most literature concerning graph based approach, the most widely used similarity measure is the cosine similarity measure. An edge then exists if the similarity weight is above some predefined threshold. Figure 3 shows an example graph for multi document.
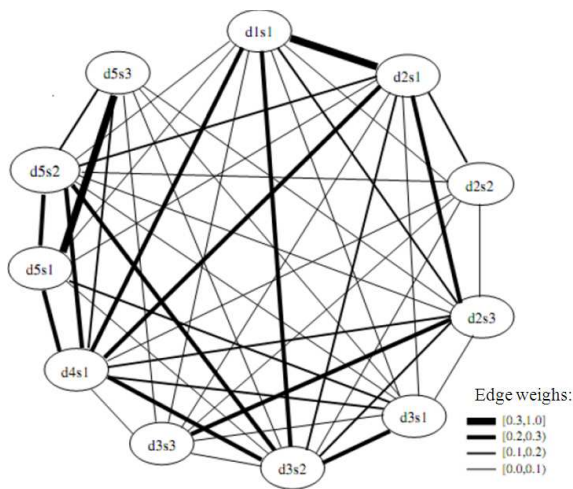
Fig. 3: Example graph (As depicted in (Erkan and Radev, 2004b). For multi document each node represents a sentence

Once the graph is constructed for a set of documents, important sentences will then be identified. It follows the idea that a sentence is considered important if it is strongly connected to many other sentences (Erkan and Radev, 2004b).

This approach differs from the cluster based approach where sentences are ranked based on its closeness to cluster centroid. Two well known graph based ranking algorithms are the HITS algorithm (Kleinberg, 1999) and the Google's PageRank (Brin and Page, 1998). Both methods have been traditionally used in Web-link analysis and social netstudys. Lexrank (Erkan and Radev, 2004b) and TextRank (Mihalcea and Tarau, 2004) are two successful graph-based ranking systems that implements these algorithms.

Further studies have been carried to make improvement through modification in the ranking algorithm. Wan and Yang (2006) assigned different weights to intra-document links and inter-document links. They give more priority to sentence with high inter-document links. In the study by Hariharan and Srinivasan (2009), they approach the graph based method differently i.e., by discounting the already selected sentence by removing the sentence from further consideration when ranking the remaining sentences.

Apart from sentence level information, Wan (2008) and Wei *et al*. (2010) devise the document-sensitive graph model to explore document impact on the graph-based summarization, by incorporating both the document-level information and the sentence-to-document relationship in the graph-based ranking process. The document-level relations are used to adjust the weights of the vertices and the strength of the edges in the graph.

The approach to graph based methods have resulted in positive feedback from the multi document summarization research communities as it was able to identify 'prestigious' sentences across the documents. The resulting graph is also able to capture distinct topics from unconnected sub-graphs. However since this approach depends heavily on sentence similarity to generate graph, it only treats sentence as bag of words without "understanding" the text. This would result the final summary to be not complete enough specifically for an informative summary generation. We will discuss further on this issue later in our proposed component based approach.

Knowledge based method: Most documents or articles will have its content related to a particular topic or event. These topics or events generally belong to a particular domain and each domain normally has its own common knowledge structure. Thus, there have been efforts made by researchers to utilize the background knowledge (i.e., ontology) to improve summarization results. In fact, many other applications have tailored their model to be ontology-driven (Shareha *et al*., 2009, Nasir and Noor, 2011).

Ontology, equipped with concise concepts and rich domain–related information, can capture the hidden semantic information. With the support of the ontology, information can be related with each other through the shared and common understanding of a domain (Khelif *et al*., 2007).

Li *et al*. (2010) developed the Ontology-enriched Multi-Document Summarization (OMS) system to generate query-relevant summary from a collection of documents. OMS first links the sentences from documents onto a domain-related ontology, then maps the given query to a specific node in the ontology and finally extracts the summary from the sentences in the sub-tree rooted at the corresponding query node. Another example is the utilization of knowledge from UMLS (www.nlm.nih.gov/research/umls/), a medical oncology, to summarize biomedical documents (Verma *et al*., 2007). Here, the authors apply the medical ontology as dictionaries of valid concepts and choose sentences that contain only those words corresponding to concepts in the ontology. Kogilavani and Balasubramanie (2009) also utilized UMLS but as an alternative they used the ontology to expand user's natural language queries with synonyms and semantically related concepts.

In previous related study, Wu and Liu (2003) manually constructed a domain specific ontology for business news articles. They determine the main

subtopics of the articles of interest by comparing the sentence words to the ontology. Sentences which are most "close" to the subtopics are then selected. A similar idea but with additional ontology features were proposed by Hennig *et al*. (2008) for sentence scoring. The features they used were tag overlap, subtree depth and subtree count.

Ontology can be useful for domain specific documents where key concepts pertaining to the domain can be identified. In most ontology based text summarization, the ontology functions to bootstrap the process of sentence selection by picking sentences which contain predefined ontology concepts. However it can be seen that the ontology is mainly used for similarity measure and as dictionary of valid concepts. One of the major concerns in ontology based summarization is the availability of the ontology itself. That is, this approach is only feasible when the ontology is available. Due to this reason, in most cases the ontology is manually constructed by experts. Unlike the previous three methods which can be applied to any domain, current ontology based method depends on a particular domain where the ontology design requires input from the domain experts.

Proposed component based approach: If we look back at previous approaches described in this study, we can observe that those approaches were mainly based on flat text feature representations without any attempts to "understand" the text. Moreover, until now, most text summarization models incorporate only bag of words as text representation and do not include much contextual information. We believe that providing comprehensive contextual information coverage would be ideal for summary creation.

In the context of news documents (which will be our research focus), different news sources reporting on a particular event tend to contain common components that make up the main story of the news. The common components of a news article consist of WHO, WHEN, WHERE, WHAT and HOW. For example, new articles on natural disaster events often contains components such as information about person, location, description of the disaster, the damages to human and properties, the relief efforts, organizations involved, the disruption of services and etc. Such occurrence of components with its information content description is what the readers usually search for while reading a news story. Moreover, as far as multi document is concerned, these components usually overlap each other and often

reappears in different parts of the document as well as in documents from other news sources. We aim to capture these components' content information to better represent multiple document coverage of a news story.

Our proposed approach will incorporate ontology learning as part of our effort to learn the relationship linking certain pairs of components content. This is relevant to the understanding of the text. In contrast, existing ontology based methods discussed in the literature merely used ontology to identify important concepts in documents. For example, Wu and Liu (2003) perform term based mapping of sentences to ontology to find the most informative concepts in a document, while Hennig *et al*. (2008) classify sentence to nodes on the ontology to identify the main topics in a document. Their efforts were mainly focused on matching the ontology concepts which appear in the text, so that frequent occurring concepts can be labeled as important topics. As opposed to their approach, we do not use ontology for identifying important concepts or topics; instead we use the idea of ontology learning to capture the relations that exist among predefined components in the news documents. Ontology learning can also be used to generate terms which are relevant to the underlying domain, in order to capture only relations among the generated terms.

This way of utilizing component's content knowledge will benefit the summarization process in two ways. First, breaking documents down to their components and capturing those links between them would produce broad information coverage for the summary, thus generating an informative summary. The other benefit is that this approach gives an "intuitive thought" on the kind of information we know that is essential to be included in the summary. The latter is close to the way how humans prepare a news related summary.

CONCLUSION

This study provides a general survey on multi document summarization approaches. Indeed, this study has been tailored in a way that researchers whom are new to the area of text summarization can grasp the idea of various multi document summarization approaches. Four types of approaches have been discussed, namely the feature based method, cluster based method, graph based method and knowledge based method. It appears that each of these methods possess its own advantages towards multi document summarization. At the same time, there are some issues or limitations pertaining to those methods. For future

improvement, we propose a novel approach, taking into account the generic components of a news story in order to generate a better summary which is well suited for an informative type summary generation. We belief that the proposed component based approach can alleviate some of the aforementioned limitations.

ACKNOWLEDGEMENT

REFERENCES

Aliguliyev, R.M., 2010. Clustering techniques and discrete particle swarm optimization algorithm for multi-document summarization. Comput. Intell., 26: 420-448. DOI: 10.1111/j.1467-8640.2010.00365.x

Binwahlan, M.S., N. Salim and L. Suanmali, 2009. Swarm based text summarization. Proceeding of the International Spring Conference on Computer Science and Information Technology, Apr. 17-20, IEEE Xploor, Singapore, pp: 145-150. DOI: 10.1109/IACSIT-SC.2009.61

Bossard, A. and C. Rodrigues, 2011. Combining a multi-document update summarization system-cbseas-with a genetic algorithm. University´e Paris.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hypertextual web search engine. Comput. Networks ISDN Syst., 30: 107-117. DOI: 10.1016/S0169-7552(98)00110-X

Dehkordi, P.K., H. Khosravi and F. Kumarci, 2009. Text summarization based on genetic programming. Int. J. Computing ICT Res., 3: 57-64.
Erkan, G. and D.R. Radev, 2004a. LexPageRank: Prestige in multi-document text summarization. University of Michigan.

Erkan, G. and D.R. Radev, 2004b. LexRank: Graph-based lexical centrality as salience in text summarization. J. Artifi. Intelli. Res., 22: 457-479.

Fattah, M.A. and F. Ren, 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Comput. Speech Language, 23: 126-144. DOI: 10.1016/j.csl.2008.04.002

Ganesan, K., C. Zhai and J. Han, 2010. Opinosis: A graph-based approach to abstractive summarization of highly redundant opinions. Proceedings of the 23rd International Conference on Computational Linguistics, (ICCL '10) Association for Computational Linguistics Stroudsburg, USA., pp: 1408.

Gupta, V. and G.S. Lehal, 2010. A survey of text summarization extractive techniques. J. Emerging Technol. Web Intell., 2: 258-268. DOI: 10.4304/jetwi.2.3.258-268

Harabagiu, S.M. and F. Lacatusu, 2010. Using topic themes for multi-document summarization. J. ACM Trans. Inform. Syst., 28: 47-47. DOI: 10.1145/1777432.1777436

Hariharan, S. and R. Srinivasan, 2009. Studies on graph based approaches for single and multi document summarizations. Int. J. Comput. Theory Eng, 1: 1793-8201.

Hariharan, S., 2010. Multi document summarization by combinational approach. Int. J. Comput. Cognition, 8: 68-74.

Hennig, L., W. Umbrath and R. Wetzker, 2008. An ontology-based approach to text summarization. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Dec. 9-12, IEEE Computer Society, USA., pp: 294. ISBN: 978-0-7695-3496-1

Jain, A.K., M.N. Murty and P.J. Flynn, 1999. Data clustering: A review. ACM Comput. Sur., 31: 264-323.

Khelif, K., R.D. Kuntz and P. Barbry, 2007. An ontology-based approach to support text mining and information retrieval in the biological domain. J. Univ. Comput. Sci., 13: 1881-1907.

Kleinberg, J.M., 1999. Authoritative sources in a hyperlinked environment. J. ACM, 46: 604-632. DOI: 10.1145/324133.324140

Kogilavani, A.A. and B.P. Balasubramanie, 2009. Ontology enhanced clustering based summarization of medical documents. Int. J. Recent Trends Engin., 1: 546-549.

Li, L., D. Wang, C. Shen and T. Li, 2010. Ontology-enriched multi-document summarization in disaster management. ProceedingS of the 33rd international ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, ACM, New York, USA., pp: 820. ISBN: 978-1-4503-0153-4

Balaji, M. and V. Kamaraj, 2011. Particle swarm optimization approach for optimal design of switched reluctance machine. Am. J. Applied Sci., 8: 374-381. DOI: 10.3844/ajassp.2011.374.381

Mihalcea, R. and P. Tarau, 2004. TextRank: Bringing Order into Text. University of North Texas.

Osborne, M., 2002. Using maximum entropy for sentence extraction. Proceedings of the ACL 2002 Workshop on Automatic Summarization, (WAS '02), Association for Computational Linguistics Stroudsburg, PA, USA., pp: 8. DOI: 10.3115/1118162.1118163

Radev, D.R., E. Hovy and K. McKeown, 2002. Introduction to the Special Issue on Summarization. J. Comput. Linguistics., 28: 399-408. DOI: 10.1162/089120102762671927

Radev, D.R., H. Jing, M. Sty and D. Tam, 2004. Centroid-based summarization of multiple documents. Inf. Process. Manage., 40: 919-938. DOI: 10.1016/j.ipm.2003.10.006

Shareha, A.A.A., M. Rajeswari and D. Ramachandram, 2009. Multimodal integration (image and text) using ontology alignment. Am. J. Applied Sci., 6: 1217-1224.

Nacy, S.M.N., S.S. Hassan and H.S. Bakhy, 2009. Geometric optimization of three-phalanx prosthesis underactuated fingers using particles swarm algorithm. Am. J. Eng. Applied Sci., 2: 381-387.

Suanmali, L., N. Salim and M.S. Binwahlan, 2009. Fuzzy Logic Based Method for Improving Text Summarization. Int. J. Comput. Sci. Inform. Sec., 2: 6-6.

Suanmali, L., N. Salim and M.S. Binwahlan, 2011. Genetic algorithm based sentence extraction for text summarization. Int. J. Innovative Comput., 1: 1-22.

Nasir, S.A.M. and N.L.M. Noor, 2011. Automating the mapping process of traditional malay textile knowledge model with the core ontology. Am. J. Econ. Bus. Admin., 3: 191-196.

Verma, R., P. Chen and W. Lu, 2007. A Semantic free-text summarization system using ontology knowledge. University of Houston-Downtown.

Wan, X. and J. Yang, 2006. Improved affinity graph based multi-document summarization. Proceedings of the Human Language Technology Conference of the (NAACL '06), Association for Computational Linguistics Stroudsburg, USA., pp: 212.

Wan, X. and J. Yang, 2008. Multi-Document Summarization Using Cluster-Based Link Analysis. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, ACM, New York, USA., pp: 906. ISBN: 978-1-60558-164-4

Wan, X., 2008. An exploration of document impact on graph-based multi-document summarization. Proceedings of the Conference on Empirical Methods in Natural Language Processing, (EMNLP '08), Association for Computational Linguistics Stroudsburg, pp: 1129.

Wei, F., W. Li, Q. Lu and Y. He, 2010. A document-sensitive graph model for multi-document summarization. Knowl. Inform. Syst., 22: 245-259. DOI: 10.1007/s10115-009-0194-2

Wu, C.W and C.L. Liu, 2003. Ontology-based Text Summarization for Business News Articles. Proceedings of the ISCA 18th International Conference on Computers and their Applications, Mar. 26-28, Honolulu, Hawaii, USA., pp: 4.

Xia, Y., Y. Zhang and J. Yao, 2011. Co-clustering sentences and terms for multi-document summarization. Comput. Linguistics Intell. Text Proc., 6609: 339-352. DOI: 10.1007/978-3-642-19437-5_28

Yu, N., D. Ji, L. Yang, Z. Niu and T. He, 2006. Multi-document summarization using a clustering-based hybrid strategy. Inform. Retrieval Technol., 4182: 608-614. DOI: 10.1007/11880592_53