A Survey on Multi-Document Summarization Approaches

Smita Bachal[1], **S.M. Sangve**[2]

[1]Department of Computer Engineering, Dnyanganga College of Engg. & Research, Pune, India

[2] Professor, Department of Computer Engineering, Dnyanganga College of Engg. & Research, Pune, India

Abstract: ***With the issue of expanded web assets and the tremendous measure of data extraction, the need of having automatic summarization systems showed up. Since summarization is required the most at present searching data on the web, where the user goes for a certain space of enthusiasm as per his query, area based summaries would serve the best. Ontology based summarization system for is presented. Summarization can be of distinctive nature extending from demonstrative summary that distinguishes the subjects of the documents to informative summary which is intended to speak to the brief depiction of the first record, giving a thought of what the entire content of record is about. This paper represents survey of recent approaches of summarization methods. We investigate approaches for multi-document summarization. Knowledge based and machine learning routines for picking the most significant sentences from reports concerning a given query are considered. In multi document summary, the general thorough quality in showing enlightening synopsis regularly needs. It is discovered that the majority of the current systems have a tendency to concentrate on sentence scoring and less attention is given to the relevant data content in various documents.***

Keywords: Summarization, machine learning, Ontology, Knowledge based, Feature based.

1. Introduction

The need of summarization has as of late expanded because of the expansion of data on the Internet. With the accessibility and pace of web, data extraction from online documents has been speed down. On the other hand, it is not simple for users to manually summarize those huge online records. In case like, when a user hunt down data about earthquake which happened in Sendai, Japan, the user will most likely get tremendous articles identified with that occasion [1]. The user would unquestionably choose summary that could outline those articles. The objective of multi-document summarization is gathering the source content into a shorter form protecting its data and general significance. The target and methodology of summarization of documents clarify the sort of summary that is created.

Approach towards summarization can be either extractive or abstractive (Radev et al., 2002). In extractive summarization, critical sentences are recognized and straightforwardly extracted from the original document, i.e. the last summary comprises of unique sentences. Then again, in abstractive summarization (Ganesan et al., 2010) the sentences which are chosen from the original report are further handled to rebuild them in the recent past linking them into last summary. This methodology generally includes profound natural language processing and sentence compression. By understanding the kind of summary i.e., indicative, informative, extractive and abstractive, we can then apply them to either single document or multi document [1][2].

This study focuses on informative and extractive type multi document summarization. The different qualities that make multi document summarization rather diverse from single document summary is that multi document summarization includes document summarization problem involves multiple sources of information that overlap and supplement. So the key work are not just distinguishing and adapting to repetition crosswise over records, additionally guaranteeing that the last synopsis is both coherent and complete.

The remaining of this study can be categorized as: We examine the four remarkable methodologies of multi document summarization and present it with related work from literature. The profits and demerits concerning these methodologies are additionally discussed. The rest of the study is sorted out as takes after: First we display the review on four multi document summarization approaches to be specific the feature based technique, cluster based strategy, graph based system and knowledge based system. At that point we point the proposed multi document summarization strategy; i.e., the component based system. At last we end with conclusion.

2. Related Work

A number of exploration study have tended to multi document summarization in the research world (Erkan and Radev, 2004a, Wan and Yang, 2008, Haribagiu and Lacatusu, 2010) also shows distinctive sorts of methodologies and accessible systems for multi document summarization.

A. Approaches of Multi-document summarization

In this study we guide our focus remarkably on four well known methodologies to multi document summarization. Our survey will be focused around the accompanying example: For every strategy, we will first talk about its primary thought. Following that, we will take some research study from related literary works [3][4][5].

B. Feature Based Method:

Extractive summarization includes distinguishing the most significant sentences from the content and set up them together to make a concise summary. At the present time recognizing critical sentences, feature impacting the importance of sentences are decided. Here we show a portion of the regular feature that has been considered for sentence choice.

- **Word Frequency**

The thought of utilizing word frequency is that essential words seem commonly in the document. The most well-known measure broadly used to compute the saying recurrence is tf and idf.

- **Sentence location**

Important data in a report is frequently secured by authors at the starting of the article. Therefore the starting sentences are expected to contain the most imperative substance.

- **Title / headline word**

Occurrence of words from the report title in sentence demonstrates that the sentence is exceedingly important to the document.

- **Sentence length:**

Very short sentences are more often than not excluded in summary as they contain less data. Long sentences are likewise not suitable to make the summary.

- **Cue word**

There are sure words in a sentence which demonstrate that the sentence is convey a useful message in the document (e.g., "essentially", "in conclusion").

- **Proper Noun**

Sentences containing proper noun, place or thing speaking to a special element suchlike name of an individual, association or place are considered vital to the record.

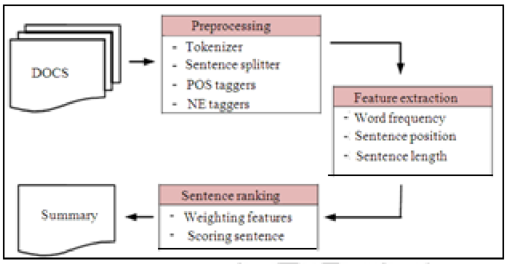

Figure 1: Feature based summarization

Above fig.1 shows feature based summarization. In any case not all text features are treated with same level of significance as a percentage of the feature have more significance or weight and some have less. Hence emphasis can be given to managing the content features focused around their vitality. This issue can be overcome by utilizing weight learning technique. Numerous researchers have been utilizing different weight learning systems in their study. Binwahlan et al. (2009) presented a novel text summarization model based on swarm brainpower procedure known as Particle Swarm Optimization (PSO) [6].

C. Cluster Based Method

The thought of clustering is to cluster comparable documents into their classes. The extent that multi reports are concerned, these objects refers to sentences and the classes speak to the cluster that a sentence has a place with. Typically, clustering algorithm can be arranged as agglomerative or partitional (Jain et al., 1999). In agglomerative clustering (otherwise called "bottom up" methodology), each one sentence is at first considered a different cluster by its own. Radev et al. (2004) spearheaded the utilization of cluster centroids for their multi-document summarizer [7][8]. Centroids are the top positioning tf-idf that speaks to the cluster. These cluster centroids are then used to recognize the sentences in each one cluster that are most like the centroid. Subsequently, the summarizer produces sentence which are most relevant to each one cluster.

D. Graph Based Method

The crucial hypothesis of graph representation is the association or connecting between items. These connections exist focused around their underlying connection. On account of content documents, the underlying connection is generally the closeness between objects for this situation, sentences. As in most wok concerning graph based methodology, the most broadly utilized comparability measure is the cosine measure. An edge at that point exists if the similarity weight is over some predefined threshold [9]. Once the graph is developed for a situated of archives, useful sentences will then be distinguished. It takes after the thought that a sentence is considered essential on the off chance that it is emphatically associated with numerous other sentences (Erkan and Radev, 2004b).

E. Knowledge Based Method

Most records or articles will have its method identified with a specific point or occasion. These themes or occasions by and large fit in with a specific area and every space ordinarily has its own basic learning structure. Consequently, there have been endeavours made via analysts to use the foundation learning (i.e., ontology) to make summarization results. Indeed, numerous different applications have customized their model to be ontology driven (Shareha et al., 2009, Nasir and Noor, 2011).li et al. (2010) created the Ontology-enriched Multi-Document Summarization (OMS) framework to produce query important summary from a gathering of documents [10][11][12]. In past related study, Wu and Liu (2003) physically developed an area particular ontology for business news articles. A comparative thought however with extra ontology peculiarities were proposed by Hennig et al. (2008) for sentence scoring [13][14].

3. Conclusion

This study gives a general overview on multi document summarization approaches. Undoubtedly, this study has been custom-made in a manner that scientists whom are new to the region of summarization can get a handle on the thought of different multi document summarization approaches. Four sorts of methodologies have been talked about, in particular the feature based system, cluster based strategy, graph based system and information based strategy. It gives the idea about each of these routines has its own particular favorable circumstances towards multi document summarization. In the meantime, there are a few issues alternately restrictions relating to those routines. For future work, a novel methodology taking into account the generic component of news story in request to create a summary which is appropriate for an informative type summarization era. We conviction that the component based methodology can reduce a percentage of the previously stated limits.

References

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

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

[8] 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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Yogan Jaya Kumar and Naomie Salim, "Automatic Multi Document Summarization Approaches", Journal of Computer Science 8 (1): 133-140, 2012 ISSN 1549-3636, 2012 Science Publications.