International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 1 (2016), pp. 1-15

http://www.ripublication.com

Ontology and Query-Focused Multi-Document Summarization System

Mr. K. Yogeswara Rao and Dr. P. V. Nageswara Rao

Department of Computer Science and Engineering,
GITAM University, Visakhapatnam, Andhra Pradesh, India
yogiindusisu@gmail.com, nagesh@gitam.in

Abstract

Due to the increasing growth of online information on the specific topic, Multiple Document Summarization (MDS) has become a non-trivial task. The MDS facilitates the user to understand the large volume of information in a short time by creating a concise and comprehensive summary. In addition, user's query based MDS system provides a consistent summary, including the core of the information. The conventional summarization techniques focus on the dynamic query based summary generation. However, it lacks in providing the entire user's information in a single convenient summary according to the particular topic. As a result, it leads to the complexity of the numerous summary generation process to each query. Hence, ensuring the effective query relevant information in extractive summary is a crucial task in MDS system. To address this constraint, this paper introduces oNtology and query focused muLti documEnt sUmmarization System (NUCLEUS). It incorporates the two essential steps such as query based summary type detection and summary generation. In the first step, NUCLEUS analyzes the document set as well as queries using ontology and Web Search Query Log (WSQL) to determine the summary type. To identify the proper context of the summary, it categorizes the document sentences based on the entities of the words in a sentence. In the second step, the NUCLEUS generates the score to each relevant query sentence using the Vector Space Model (VSM) and then, the sentences are compressed by linguistic structure analysis. Eventually, it measures the edge weight between the sentences to order coherently the sentences which having high salience and information diversity in the final summary. The evaluation results show that the NUCLEUS system can obtain significant improvement over the conventional summarization method.

Keywords: Multi-document summarization, Query-focused, sentence score, WSQL, and ontology

1. Introduction

Due to the overwhelming amount of information on the web documents, an effective automatic text summarization is a vital process to improve the user convenient summary. The automatic text summarization [1] is the process of condensing a document or document set into an information-rich summary. In recent years, automatic Multi-Document Summarization (MDS) has tremendous attention in both business and research fields. The main goal of MDS system is to generate the summary, delivering a concise, consistent, and comprehensive information content from the document collections. MDS incorporates the multiple sources of documents that contain the similar text information on the same topic [2]. Hence, MDS is to focus on recognizing the originality of the document set to generate the final summary with both coherent and thorough information, and identifying and extracting the redundant sentences. However, the existing MDS system faces the cogent sentences identification problem while extracting the sentences [3].

An Ontology is one of the rich sources to identify the core information about the texts based on the semantic representation of the hierarchical structure. To accurately generate the topic-based summary, the search query log is another one of the rich sources to understand the valuable information of user's search strategies and preferences. With improved performance of query-based summarization, MDS system requires the pre-given queries to generate the cogent summary from the multiple document sentences. The previous researchers initially treat the sentence ranking problem in query based MDS system and then focus on the coverage, redundancy removal, and coherence during sentence selection [4, 5]. Nevertheless, the exact query, i.e., user's search logs relevant summary generation is still a major constraint in MDS system. Hence, the proposed approach contemplates the frequent queries from the set of pre-given queries to identify the most relevant sentences in the document set.

The main contribution of oNtology and qUery foCused muLti documEnt sUmmarization System (NUCLEUS) includes two primary processes such as identifying the summary type using ontology and Web Search Query Log (WSQL) and generating query focused summary.

- The NUCLEUS approach investigates the succinct and informative summary generation based on the information from the user's queries and the multiple documents on the specific topic.
- NUCLEUS approach analyzes the documents by recognizing the entities of the keywords appearing in the sentences of the document using ontology. After recognizing the appropriate entities, it categorizes the sentences according to the entity types in a sentence. It analyzes the user's queries that are retrieved from WSQL based on the sentence categories to discover the summary type.
- It generates the score to query relevant document sentences based on the query as well as a document set and compresses the sentences using linguistic structure analysis. Finally, it constructs the query focused summary based on the edge weight measurement while ensuring the information diversity, salience, and coherence.
- The experimental results show that the NUCLEUS approach significantly

improves the recall of the final summary than the conventional summarization method.

1.1 Problem statement

The extractive summarization methods face the two issues such as sentence ranking and sentence selection while ensuring the limited length of the summary. The sentence ranking and sentence selection require several factors like sentence relevance to a query, redundancy, salience, and coverage. In MDS system, the repetition of information is considered as the most significant constraint, since multiple documents discuss the same topic of information. The MDS incorporates the multiple sources of the same information. Hence, it creates the content overlapping issues and sentence selection complexity during summary generation. The existing research works depend on the sentence similarity that considers the bag of words representation of the sentences. It does not take into account the original context of the text when there is a link to each text with others in a sentence. Hence, this summary generation is unable to provide the informative content due to the lack of understanding the text. The query-focused summarization additionally provides an informative final summary instead of generating the summary without considering the query relevance. However, the query-focused MDS faces two critical issues that are lacking in highly query relevant summary generation, and measurement of salience and diversity for a batch of sentences. To aggregate all the informative sentences from the multiple documents into a single summary is a challenging process when the summary has a limited length or capacity. Hence, the proposed approach exploits the user's previous queries to build the summary with the coverage of significant information on the specific topic.

2. Related work

In this section, this paper discusses the several existing summarization techniques in MDS system. The Sentence-clustering based extractive summarization method [6] discusses the determination of latent topic sections and information-rich sentences. An extractive multi-document summarization employs Generic Relation Extraction (GRE) for identifying relation that improves the performance without modifying the model parameters [7]. The Multi-document Rhetorical structure (MRS) [8] shows the multiple relationships of a variety of text units with different levels based on the rhetorical, semantic, and temporal relationships. An extractive MDS method obtains the desired summary based on the hierarchical topic model of Latent Dirichlet Allocation (hLDA) and sentence compression. This hierarchy model and compression provide the semantic analysis based concise summary [9]. The Link analysis based summarization approach [10] employs the rhetorical relation between the sentences to reduce the redundancy that is obtained by Support Vector Machines (SVMs). An approach [11] discovers the locally relevant sentences in each document rather than identifying the whole document set and then it constructs the summary with high quality. The Sentence extraction approach employs SVM classifier to discover the semantically relevant sentences using ontology [12]. The Exploiting Ontological Relations for Automatic Semantic Tag Recommendation (ATR) approach [13]

focuses on the problem of semantically related information by exploiting ontology and semantic tags. It requires the tagged information to summarize the content-related information about the documents.

2.1 Query based extractive summarization

Hiersum [14] employs the combination of KL-divergence criterion to select the relevant sentences based on the subtopics identification, is the probabilistic generative model of LDA based topic model. An approach [15] extends the factorial LDA model with a together consideration of drug, effect and its causes for generating the extractive summary. The researchers in [16, 17] discusses the LDA, and the approach [18] directly utilizes hierarchical LDA (hLDA) model to discover the sentence structures of supervised summarization. The Query based summarization [19] builds the desired summary by extracting the sentences based on the correlation with the query and global connectivity. The Document Graph (DG) model based summarization technique obtains the desired summary by performing semi-supervised query-oriented summarization [20]. The Query-oriented MDS approach introduces the query-sensitive graph based sentence ranking algorithm to estimate the edge weight between the sentences for obtaining the most relevant summary [21]. An approach [22] discusses the machine learning techniques based query oriented summarization using Support Vector Regression (SVR). The Fast and accurate query based MDS (Fastsum) [23] exploits the SVR model to summarize the query-relevant sentences based on the word frequency features of clusters, documents, and topics.

Most of the current approaches develop the document content model based on the word frequency based methods, graph-based representations, structured probabilistic topic models, and linguistically motivated techniques. These techniques do not coherently build the extractive summary due to the irrelevant content of the user's queries. However, some of the researchers focus on analyzing the context of user's frequent queries on the specific topic.

3. An Overview of NUCLEUS methodology

To create the user convenient summary with fine attributes of salience and coherence, the NUCLEUS investigates the query-relevant summary generation. The conventional query-focused summarization techniques lack concentration on the most frequent questions on the particular document set since these techniques have taken into account a single query as the topic of the summary. The NUCLEUS exploits the WSQL of the specific document set to improve the information diversity by incorporating the information in the summary that are relevant to the entire queries. The NUCLEUS approach comprises two processes such as identifying the summary type using ontology and WSQL and generating the query focused summary. The overall proposed methodology process is illustrated in Fig.1.

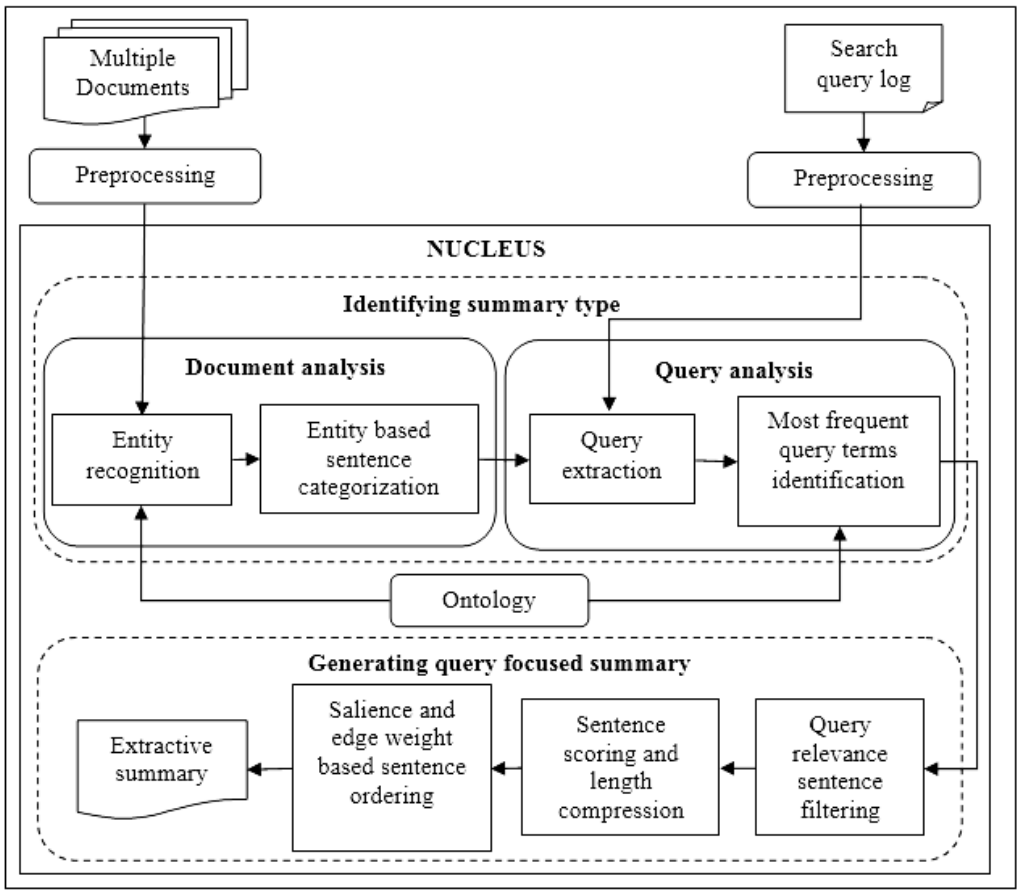

Figure 1: The NUCLEUS methodology

Identifying Summary Type using Ontology and WSQL: The NUCLEUS approach analyzes the input document set using an ontology to recognize the entity of each keyword in the document set. It extracts the queries from the web search logs that are relevant to the entities of each category in the document set. Then, it determines the frequent queries on each category, prefers that type of queries related sentences as the most important in the document set, and identifies the maximum coverage information of the summary.

Generating Query-Focused Summary: The NUCLEUS approach exploits the semantic relation between the query terms and document sentences to filter the query-relevant sentences from the document set. It generates the score for each relevant query sentence based on the recognized entities and pertinence value of the document as well as a query. It reduces the sentence length using linguistic structure analysis and measures the salience of the sentences based on the query terms. Finally, it coherently summarizes the high scored and salience sentences by ordering the edge weight based sentence pair.

3.1 Identifying Summary Type using Ontology and WSQL

To determine the summary type, the NUCLEUS performs two analysis by exploiting the ontology and WSQL. Firstly, it analyzes the document set using an ontology to extract the entity of all the keywords or terms in the document sentences and then, that sentences are grouped into a category based on the entities. Secondly, it analyzes each sentence category related queries using WSQL and query preference based on the frequent queries. It discovers the summary type, including the query-relevant information that contemplates the entire set of queries related information. Eventually, it comprises the ontology-based query terms, document sets, and summary type.

3.1.1 Analysing Document Set using Ontology

The NUCLEUS examines the input document set that includes documents, sentences, and terms. The initial stage of this document analysis is the preprocessing procedure that extracts the sentences from the multiple documents. The preprocessing procedure involves the sentence division, tokenization, Part-Of-Speech (POS) tagging, stop-word removal, and stemming. At this stage, the NUCLEUS retains the keywords of the document set in sentence vector representation that facilitates the identification of the original context of each keyword in the document sentences. The keywords of the document sentences are presentedto the ontology to recognize the entities of the keywords. The ontology provides the entity for each keyword (t) in which each keyword involves one or more entities. In the case of a keyword having more than one entity, the NUCLEUS determines the corresponding entity of each keyword by generating document context related entity score (E_S). The proposed entity score is the summation of similarity between an entity (E_n^t) and document context (d_j), and coherence between the document set (D_S) and entity.

The NUCLEUS categorizes the sentences based on the entities of the keyword presence in the document sentences. It groups the sentences under the category (C) of repetitive or similar entities when a sentence contains the repetitive entities more than the threshold value. This category segregation within a document set is beneficial to retrieve the frequent queries and to identify the frequent query relevant sentences. As a result, the document set contains different categories, and each category comprises the group of sentences.

$$W(t_i,C) = \log(ft_i,C) * \log \left(\frac{M}{mt_i} \right) \quad \text{.....(1)}$$

In equation (1), the weight of terms in document sentences can be measured. It is measured by using the category of the document sentences, i.e. i^{th} term appearance in the category based term frequency (tf) and inverse document frequency (idf) measurement. Where, ft_i,q is the term frequency in the entity based categories, M is the total number of categories, and mt_i is the number of categories which comprises i^{th} term.

3.1.2 Analysing Queries through Entity based WSQL

To capture the actual need of users, the NUCLEUS exploits the WSQL in the context of the document set regarding sentence category. The entity based sentence categorization is beneficial to recognize the relevant context queries instead of using entities of the whole document set that may distract the context of the document set. The NUCLEUS analyzes the query similar to the document analysis in which importance of the query terms is identified by measuring the term weight in the user's queries. From the WSQL, the NUCLEUS retrieves entity category based queries on each specific category. It highly prefers the frequent queries on each category and other than that are also involved in the key queries set but with low preference. The weight of query term in the queries is computed by using tf-idf measurement as similar to the one as shown in equation (1). However, it is measured using only the queries, not the sentence categories.

The importance based query selection indicates the accurate final summary generation that covers the maximum information of the frequent queries. Other than the frequent queries related information, they are partially involved in the final summary. The NUCLEUS creates the new set of query terms using ontology that provides semantically related keywords to enrich the summary generation. It employs the distance threshold level based semantic link of the keywords while avoiding the original context distraction. It attempts to cover the information with accurate and coherent aggregation of the document sentences. Finally, this phase holds the set of semantically related query terms, document sentences, and a set of keywords with corresponding entities. This query term concludes the summary type for the corresponding document set. It identifies the summary type as either a desired summary or a general summary based on the user's queries. The coverage of information, in summary, depends on the majority of the sentences in the category and the frequent queries to build the comprehensive summary.

3.2 Generating Query-Focused Summary

The NUCLEUS approach generates the concise and comprehensive final summary by applying optimal sentence scoring and sentence selection method based on the user's information need regarding a query on the specific topic. Initially, the NUCLEUS identifies the query-relevant sentences and then generates the score for each sentence using query terms as well as entity score of the keywords in a sentence. It reduces the length of the salience sentences using linguistic structure analysis. The resulting, final summary builds up by coherently selecting and ordering the optimal sentences based on the edge weight between the sentences. The proposed algorithm is shown in Alg.1.

3.2.1 Ranking and Compressing the Query Relevant Sentences

At first, the NUCLEUS determines the query-relevant sentences from the whole document sentences based on the semantic similarity score between query terms (q_t) and document sentences (S). The semantic similarity measurement is based on the Vector Space Model (VSM) in which semantic similarity represents the ontology based meta-features of the query terms. The VSM of cosine similarity measurement shows the distance between document sentence and query that facilitates the sentence

selection based on the importance of the sentence (Sim(S, Q)). To generate the succinct and informative summary, the relevance analysis is necessary which measures the relevance between the user's query and document sentences. It is used to reflect the significance of each sentence in the document set. After ensuring the sentence query relevance using VSM, the NUCLEUS approach assigns the score to each sentence in the document set. The sentence scoring or ranking procedure incorporates the entity score, frequency, position, distance of each keyword, and sentence query relevance (S, Q) in the document set.

$$R(S_j) = \frac{Sim(S,Q) * \sum_{i=1}^{n} [ES(t_i) * \log(F(t_i).P(t_i)) * (1/D(t_i))]}{|E(S_{ij})|} \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

From equation (2), the first term of Sim(S, Q) is derived from the relevance analysis, and the second summation term represents the similarity between term and document (Sim (t_i, D)). (Sim(t_i,D)) includes the frequency of the i^{th} term in each sentence category, the position of an i^{th} term in each document, and distance of i^{th} term in the ontology. $|E(S_i^j)|$ denotes the entities in a sentence and i=1,...n represents the number of keywords or terms in a sentence. The NUCLEUS approach reduces the length of the scored sentences by removing the unnecessary words from the document sentences. It determines the trivial terms by applying the linguistic structure analysis of the scored document sentences. It simplifies the burden of the sentences without fluctuating the original context of the sentences. Finally, this step contains the query-relevant high scored ($R(S_j)$) and compressed sentences ($C(S_j)$) based on the threshold level of summary length. It comprises the complete queries related information regarding document sentences without performing the proper concatenation.

3.2.2 Summarizing the High-Scored Sentences

To maintain the proper concatenation of the sentences, the NUCLEUS approach selects the sentences with high information diversity, salience, and coherence. To improve the information diversity, the NUCLEUS approach employs the semantic similarity by removing the redundant informative sentences. It improves the anti-redundancy in the final summary due to the consideration of minimum similarity score to the previously summarized sentences while selecting the optimal document sentences. After removing the redundant sentences, the NUCLEUS sorts the high scored sentences based on the salience measurement that shows the inherent relation between the sentence and the document. Also, it contemplates the salience measurement between the queries and sentences to generate comprehensive query-focused summary using similarity measurement. At this stage, this step retains the high similarity of the sentences to reflect the strong context relationship with the document as well as the corresponding queries. Furthermore, it orders the sentences to improve the accuracy regarding understanding the context easily. The scored, as well as compressed salience sentences, are ordered to maintain the coherency. The proposed sentence selection and ordering depend on the sentence-queries relationship

and edge weight ($W(e_{j1j2})$) respectively. To maintain the coherency, the measurement of edge weight between the sentences is necessary.

Input: Documents

Output: Summary

Let $t_i=\{t_1,...t_n\}$;$S_j=\{S_1,...S_m\}$;$D_k=\{d_1,....d_n\}$;$Q_i=\{q_1,...q_m\}$

While d ϵ D_k**do**

for allDocuments(D_S) **do**

//Identifying summary type

D_S<- Preprocessing

D_S -> Ontology

Recognize entity (E_n) for Preprocessed (D_S) <- Ontology

for all recognized entities **do**

Generate E_S using document and ontology

if($S_j(E_n^t)\geq\alpha$) **then**

Categorize S_j under E_n^t

endif

for allqueries(Q) **do**

Retrieve S_j^C based queries from search log

Determine frequent Q (f_Q) on specific S_j^C

if($Q=f_Q$) **then**

Prefer the maximum importance to Q

else if

Set Q as less important queries

end if

Ontology (Q) -> semantic query terms

Identify ST in terms of coverage of S_i according to query importance

Algorithm 1: The proposed algorithm of NUCLEUS

The NUCLEUS selects the sentences with the high preference of frequent queries related information on the sentence category to generate the summary. Equation (3) reflects the correlation in terms of contextual similarity between the sentences S_{j1} and S_{j2}. $tf_{Sem}(S_{j1})$ denotes the semantic term frequency of the i^{th} term in $j1^{th}$ sentence in which i=1,...n indicates the number of words or terms present in a sentence. $N(S_{j1})$ and $N(S_{j2})$ represents the number of terms present in $j1^{th}$ and $j2^{th}$ sentence respectively. The semantic relationship between the termsis identified by using the ontology. It measures the similarity of the terms between the sentences except stop-words. The NUCLEUS orders the sentences from the high similarity or edge weight of the sentences until to reach the maximum summary length (SL). The threshold based summary length is fixed in MDS. Thus, it constructs the query-focused succinct and informative summary.

4. Experimental Evaluation

To considerably illustrate the performance of the NUCLEUS approach, this paper compares the implementation results of the NUCLEUS and a baseline ontology based ATR approach [13].

4.1 Experimental setup

The NUCLEUS approach is implemented using Java platform. To recognize the tag of the terms in a sentence, it exploits Stanford parser. It employs the Yago ontology and WSQL to determine the semantic relation and entities, and query-relevant information. It uses Tregex to reduce the sentence length based on the linguistic structure analysis. The proposed implementation measures the quality of the summary based on the overlapping information or sentences in the author made catchphrases, i.e., reference summary, and the system generated final summary.

4.1.1 Dataset

The NUCLEUS approach employs the multiple documents from Query Chain Focused Summarization (QCFS) dataset [25]. It contains Asthma, Lung Cancer, Alzheimer's Disease, and Obesity related Document sets. The NUCLEUS exploits Asthma related document set that comprises the 125 documents and 1924 sentences. Query chain contains the five documents and 15 sentences that are related to the Asthma. It provides the manual summary to evaluate the performance of the proposed summary.

4.1.2 Evaluation metrics

Precision: Precision is the ratio between the number of sentences occurring in both the final summary and reference summary and a number of sentences in the final summary.

Recall: Recall is the ratio between the number of sentences occurring in both the final summary and reference summary and a number of sentences in the reference summary.

F-measure: F-measure is the composite measure of precision and recall combination.

$$\text{F-measure} = 2 * \left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

4.2 Experimental results

4.2.1 Precision

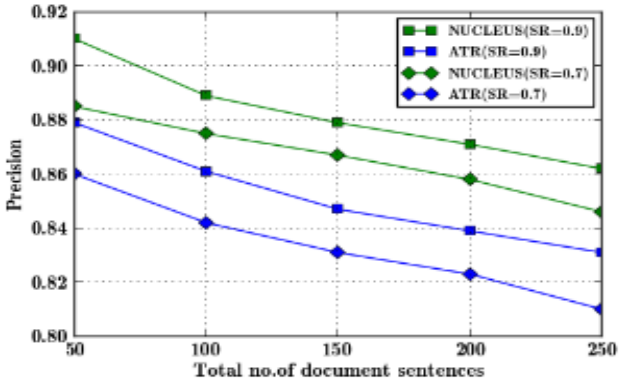

Figure 2: Precision vs. Total number of document sentences

Fig.2 demonstrates the precision of both the NUCLEUS and ATR approach while varying the total number of input document sentences as well as the Salience Rate (SR). SR is the ratio between the number of document context related sentences and the total number of document sentences. On the whole, the precision value decreases slowly with an increase in the total number of document sentences. At the point of 0.91 precision value and SR= 0.9, 200 number of document sentences are sufficient to construct the summary with the length of 15-20 sentences. Hence, its 250 sentences may degrade the performance of the summary due to the redundant sentences. When SR=0.7 and number of document sentences=250, the precision value of the NUCLEUS approach has increased by 2.9% than the ATR approach, since NUCLEUS employs the entity level based sentences instead of using semantic tag knowledge to summarize the document sentences.

4.2.2 Recall

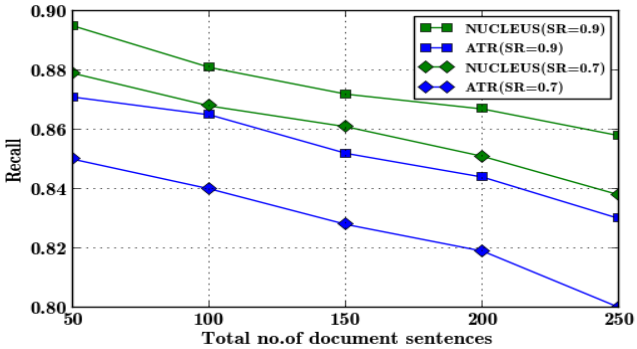

Figure 3: Recall vs. Total number of document sentences

The recall of NUCLEUS and ATR approach is comparatively shown in Fig.3. It suddenly decreases while increasing the total number of input sentences in the multiple documents. If the salience rate is high i.e. Sentences in the summary is relevant to the context of the document set, the recall value increases. When SR=0.7 the recall value of NUCLEUS approach is nearly equal to ATR approach when SR=0.9 due to the query as well as the document the context-based selection of the informative sentences. At the level of SR=0.7 and varying the number of document sentences from 50 to 250, the NUCLEUS approach decreases by 4.66% but, ATR approach suddenly decreases by 5.88% due to the maximum importance of the query and semantic link based sentence score generation.

4.2.3 F-measure

Fig.4 indicates the F-measure value of the NUCLEUS approach while varying the sentence-category matching factor and Number of Queries (NQ). NQ is the total number of input queries that are relevant to the entity based sentence category of the document set. The sentence-category matching factor is the ratio between the number of matched entities in a sentence to the category and the total number of entities in a sentence. If a sentence has five entities, four entities are matched into one category, a minimum number of queries are sufficient to generate the summary. When NQ=10, the F-measure value is gradually increased by 2.55% while sentence-category matching factor from 0.2 to 0.6. However, to increase 2.77% of F-measure value, it requires NQ=20 when sentence-category matching factor from 0.2 to 0.8. Hence, the NUCLEUS approach improves the performance due to the consideration of linguistic structure analysis based sentence length reduction to construct the effective summary.

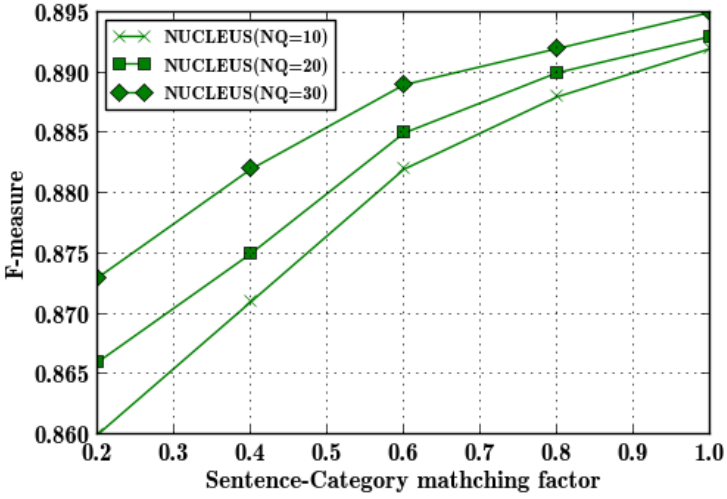

Figure 4: F-measure vs. Sentence-category matching factor

4.2.4 Anti-redundancy

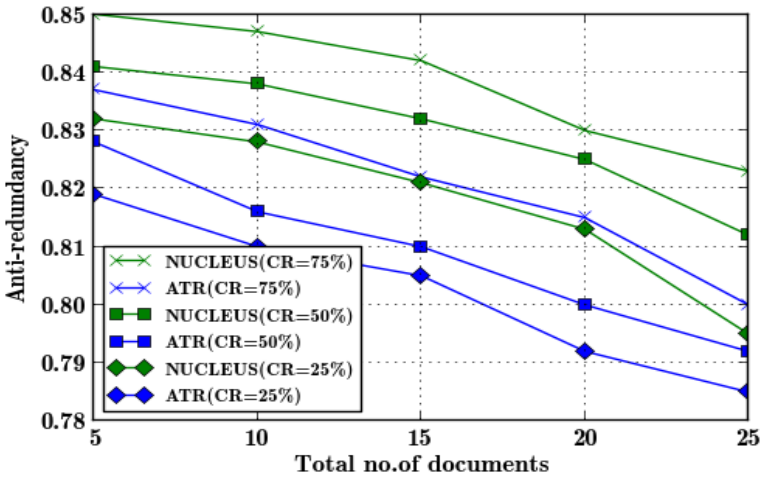

Figure 5: Anti-redundancy vs. Total number of documents

Anti-redundancy of both NUCLEUS and ATR approach is illustrated in Fig.5 while changing the total number of documents and Compaction Ratio (CR). CR is the ratio between the total number of sentences in the summary (S_{Sen}) and a total number of sentences in the input documents (D_{Sen}) i.e. CR=(1-(S_{Sen}/D_{Sen})). When increasing the total number of similar documents, the anti-redundancy value gets decreased due to the redundant, similar informative sentences. When increasing the CR from 25% to 75%, the anti-redundancy of NUCLEUS increases by 3.52% than ATR approach, since, NUCLEUS considers the together of the query and document context to identify the sentence importance. If CR=75%, NUCLEUS approach marginally decreases the anti-redundancy by 3.17% but, ATR suddenly decreases the anti-redundancy by 4.42% while increasing the number of similar documents. The NUCLEUS balances the anti-redundancy until reaching the particular level, i.e., the number of documents where in the semantically redundant informative sentences are removed using query information and document context.

5. Conclusion

This paper introduces and employs the NUCLEUS to query focused multi-document summarization that generates the coherent and intelligible summary. The main objective of this NUCLEUS approach is achieved by focusing on the ontology and entire query set on the specific topic. It jointly considers the document set as well as search queries to create the final summary with the coverage of entire convenient, informative sentences. It assigns the score to the query-relevant sentences and reduces the sentence length to obtain the consistent summary length. Finally, it constructs the summary that incorporates the query relevance, content salience, information diversity, and coherence. The evaluation results show that NUCLEUS significantly

outperforms the conventional summarization method by improving the recall of the final summary.

References

[1] Dalal, Vikram, and Latesh G. Malik, "A survey of extractive and abstractive text summarization techniques", IEEE 6th International Conference on Emerging Trends in Engineering and Technology (ICETET), pp.109-110, 2013

[2] Wan, Xiaojun, Jianwu Yang, and Jianguo Xiao, "Manifold-Ranking Based Topic-Focused Multi-Document Summarization", ACM Proceedings of the 20th international joint conference on Artifical intelligence, Vol.7, pp.2903-2908, 2007

[3] Ding Yuan, "A Survey on Multi-Document Summarization", ACM Proceedings of the HLT-NAACL 03 on Text summarizationworkshop, Association for Computational Linguistics, Vol.5, pp.65-72, 2003

[4] L. Li, K. Zhou, G. Xue, H. Zha, and Y. Yu, "Enhancing diversity, coverage and balance for summarization through structure learning", ACM Proceedings of the 18th international conference on World wide web, pp.71–80, 2009

[5] X. Li, Y. Shen, L. Du, and C. Xiong, "Exploiting novelty, coverage and balance for topic-focused multi-document summarization", ACM Proceedings of the 19th international conference on Information and knowledge management, pp.1765–1768, 2010

[6] Aliguliyev, Ramiz M, "A new sentence similarity measure and sentence based extractive technique for automatic text summarization", Elsevier transaction on Expert Systems with Applications, Vol.36, No.4, pp.7764-7772, 2009

[7] Hachey Ben, "Multi-document summarization using generic relation extraction", ACM Proceedings of the Conference on Empirical Methods in Natural Language Processing, Vol.1, pp.420-429, 2009

[8] Yong-dong, Xu, Wang Xiao-long, Liu Tao, and Xu Zhi-ming, "Multi-document summarization based on rhetorical structure: Sentence extraction and evaluation", IEEE International Conference on Systems, Man and Cybernetics, pp.3034-3039, 2007

[9] Liu, Hongyan, and Lei Li, "Multi-document summarization based on hierarchical topic model", IEEE 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp.88-91, 2011

[10] Zahri, Nik Adilah Hanin Binti, and Fumiyo Fukumoto, "Multi-document summarization using link analysis based on rhetorical relations between sentences", Springer Berlin Heidelberg on Computational Linguistics and Intelligent Text Processing, pp.328-338, 2011

[11] Villatoro-Tello, Esaú, L. Villaseor-Pineda, Manuel Montes-y-Gómez, and D. Pinto-Avendao, "Multi-Document Summarization Based on Locally Relevant Sentences", IEEE Eighth Mexican International Conference on Artificial Intelligence, pp.87-91, 2009

[12] Hennig, Leonhard, Winfried Umbrath, and Robert Wetzker, "An ontology-based approach to text summarization", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol.3, pp.291-294, 2008

[13] Alexopoulos Panos, John Pavlopoulos, Manolis Wallace, and Konstantinos Kafentzis, "Exploiting ontological relations for automatic semantic tag recommendation", ACM Proceedings of the 7th International Conference on Semantic Systems, pp.105-110, 2011

[14] Haghighi Aria, and Lucy Vanderwende, "Exploring content models for multi-document summarization", ACM Proceedings of Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp.362-370, 2009

[15] Paul Michael J., and Mark Dredze, "Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models", In HLT-NAACL, pp.168-178, 2013

[16] Li Jiwei, and Sujian Li, "Evolutionary Hierarchical Dirichlet Process for Timeline Summarization", pp.556-560, 2013

[17] Mason Rebecca, and Eugene Charniak, "Extractive multi-document summaries should explicitly not contain document-specific content", ACM Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pp.49-54, 2011

[18] Celikyilmaz Asli, and Dilek Hakkani-Tur, "A hybrid hierarchical model for multi-document summarization", ACM Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp.815-824, 2010

[19] Ye Xinghuo, and Hai Wei, "Query-Based Summarization for Search Lists", IEEE First International Workshop in Knowledge Discovery and Data Mining, pp.330-333, 2008

[20] Wang Wei, Sujian Li, Jiwei Li, Wenjie Li, and Furu Wei, "Exploring hypergraph-based semi-supervised ranking for query-oriented summarization", Elsevier transaction on Information Sciences, Vol.237, pp.271-286, 2013

[21] Wei, Furu, Yanxiang He, Wenjie Li, and Qin Lu, "A Query-Sensitive Graph-Based Sentence Ranking Algorithm for Query-Oriented Multi-document Summarization", IEEE International Symposium on Information Processing (ISIP), pp.9-13, 2008

[22] Ouyang, You, Wenjie Li, Sujian Li, and Qin Lu, "Applying regression models to query-focused multi-document summarization", Elsevier transaction on Information Processing and Management, Vol.47, No.2, pp.227-237, 2011

[23] Schilder, Frank, and Ravikumar Kondadadi, "FastSum: fast and accurate query-based multi-document summarization", ACM Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, pp.205-208, 2008

[24] http://www.cs.bgu.ac.il/~nlpproj/QCFS/dataset.html