International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-05, Aug 2018

A Review on Approaches for Information Extraction from Multiple Documents

Praveen K Wilson, Department of Information Technology, College of Engineering Perumon, praveenkwilson@gmail.com

Dr. J R Jeba, Designation, Department of Computer Application , Noorul Islam Centre For Higher Education, jrjeba@rediffmail.com

Abstract -**Text mining, a sub part of the vast area data mining is the process of extracting information from a large amount of unstructured textual resources. The process of text mining usually involves structuring the input text, deriving patterns within the structured data, and finally evaluation and interpretation of the output. As the amount of online information grows, the problem of extracting required information becomes more difficult, which leads to information overload. If this is the situation information extraction cannot be done manually and needs some computer aided approaches. Automatic text extraction is a system of extracting text by computer when a text is given as input and the output is a shorter and less redundant form of the original text. But recently information about a single topic is found in various sources such as websites, journals, newspapers, text books etc., for which multi-document extraction is required. Multi-document text extraction means to retrieve salient information about a topic from various sources. In this process, multiple textual resources are given as the input and all these inputs are processed separately. An efficient algorithm is used for analyzing all these documents and extract reliable and important information from each of the document. Then significant sentences are extracted from each output units and re-organized them to get multi-documents' summary.**

Keywords — Summarization, Information Extraction, Multi-document summary, Text Mining, Clustering, Data Abstraction

I. Introduction

With the continuing growth of World Wide Web online information content is gigantic day by day. Hence it is increasingly important to provide improved mechanisms to find and express information effectively. Traditional methods did not use any semantic approaches but rank documents based on maximum relevance to the user query. Some systems also include sub-document relevance assessments and convey this information to the user. More recently, single document summarization systems provide an automated generic abstract or a query relevant summary. However, large scale information retrieval and document summarization that have implemented are not complete or accurate when we approach semantically. Among these two areas summarization facing greater challenges when we intending to implement it as a semantic model.

II. Multi-document Summarization - Approaches

As per general definition, text summarization is a process of extracting summary of a document without losing its knowledge content but reducing the word count which may helps to reduce reading time. If we do it manually, an efficient linguist can produce a better output, but may take much time for reading, and producing summary. But we are in a day of information explosion. Day by day millions of documents are added to the www without any restrictions hence manual summarization is a time consuming job in this era. So we have to think about some automatic text processing methodologies which may generate the summary with accuracy like a linguist but in a time bound manner. This summarization has two approaches, it may be extractive or may be abstractive. In extractive approach it selects some of the sentences from the parent document then shows it as a summary. But this cannot be considered as a good summary. Better one is an abstractive approach in which the summary does not extracting some sentences but, it abstract the information from the document. Various approaches exist but majority follows the extraction approaches hence those are not technically perfect. Here we are discussing some approaches, its specialties, advantages and disadvantages. Automatic text summarization produces a concise summary by following abstraction or extraction of input text using techniques such as statistical , linguistical or combination of the two [1].

Since the techniques used in single document

DOI : 10.18231/2454-9150.2018.0607

summarization can also be used in multi-document summarization, one may think that the approaches are similar. But as per Jade we can identify at least some significant differences [2]:

In case of a single document we may not bother much about it redundancy, since the chance is very low. But in the case of multi-document summary we have lots of documents which may share same information content as input. Hence special care may take to eliminate redundancy. Another issue is the amount of input document. It is much larger compared to single document summary. Co-reference is another challenge to solve for getting a better summary.

Multi-document summarization has drawn much attention in recent years and it exhibits the practicability in document management and search systems. Multi-document summary can be used to concisely describe the information contained in a cluster of documents and facilitate the users to understand the document cluster. For example, a number of news services, such as Google News1, NewsBlaster2, have been developed to group news articles into news topics, and then produce a short summary for each news topic. The users can easily understand the topic they have interest in by taking a look at the short summary.

III. Existing Techniques – A Comparison

Research on single document summarization has turned into a more complicated era of multi-document summarization. Since the research ages more than twenty five years different approaches are available for retrieval of information from different sources. A brief description of various techniques on multi-document summarization is given below:

i) NeATS: - Chin-Yew Lin and Eduard Hovy[7]

It describes an extraction-based multi-document summarization system NeATS. It follows a three stage approach in which the stages are: content selection, filtering, and presentation. The strategies employed by NeATS are worked to some extent. One notable advantage is it is simple since it uses existing techniques. But drawbacks are many; It computes the likelihood ratio λ to identify key concepts without resolving pronouns. Its not demanding any new technology. It is only an extraction based approach. It doesn't use the concept of abstraction for further compression.

ii) Based on Two-Level Sparse Representation Model - He Liu, Hongliang Yu, Zhi-Hong Deng[8]

It is also a multi-document summarization approach which follows sparse coding technique. It develops a two level sparse model to generate the summary. It is an extraction based approach, which generates summary by extracting some of the sentences from different input documents. Here it ensure the quality of the summary by following three measurable properties such as Coverage, Sparsity and Diversity. One specialty of this approach is that we can recreate original document from its summary. Disadvantage is that the concept of abstraction not present which helps to compress the document more.

iii) Maximizing Informative Content-Words - Wen-tau Yih Joshua Goodman Lucy Vanderwende Hisami Suzuki[9]

This system has two components: one component uses machine learning to compute scores for each word in the set of documents to be summarized (which is called the "document cluster"); the other component uses a search algorithm to find the best set of sentences from the document cluster for maximizing the scores. The summary is then generated through a simple greedy search algorithm: it iteratively selects the sentence with the highest-scoring content-word, breaking ties by using the average score of the sentences. This continues until the maximum summary length has been reached. Positive things are it is relatively simple to implement and features like frequency, positions are considered which improves accuracy. Negatives are here concept identification based on frequency, and position only. Coherence between sentences in the summary is not ensured.

iv) Topic-Focused Multi-Document Summarization - Xiaojun Wan, Jianwu Yang and Jianguo Xiao[10]

The manifold- ranking process can naturally make full use of both the relationships among all the sentences in the documents and the relationships between the given topic and the sentences. The ranking score is obtained for each sentence in the manifold-ranking process to denote the biased information richness of the sentence. Then the greedy algorithm is employed to impose diversity penalty on each sentence. The summary is produced by choosing the sentences with both high biased information richness and high information novelty. Advantage of this technique is summary generated based on the topic. And it's an extractive method hence comparatively simple. Disadvantage is it needs some additional information to generate summary. Here existing techniques are used. Ie, manifold ranking, greedy algorithm etc. No abstraction is used in this approach.

v) Using Sentence Extraction - Jade Goldstein* Vibhu Mittal t Jaime Carbonell* Mark Kantrowitzt[11]

This paper presented a statistical method of generating extraction based multi-document summaries. First segment the documents into passages, and index them using inverted indices, then identify the passages relevant to the query using cosine similarity. Then a statistical algorithm is used to find relevant passages and hence make a summary. This approach is simple and easy to implement. Drawback is even it uses natural language approach, coreference is not

DOI : 10.18231/2454-9150.2018.0607

resolved. Sentences in the summary may be disjoint and semantic treatment is not done here.

vi) Using an Approximate Oracle Score – John M. Conroy, Judith D. Schlesinger, Dianne P. O'Leary[12]

It introduces an oracle score based upon the simple model of the probability that a human will choose to include a term in a summary. The oracle score demonstrated that for task-based summarization, extract summaries score as well as human-generated abstracts using ROUGE. It then demonstrated that an approximation of the oracle score based upon query terms and signature terms gives rise to an automatic method of summarization. Advantages of this approach is redundant sentences can be removed. The procedure is comparitively simple one.

vii) Coherent Multi-Document Summarization - Janara Christensen, Mausam, Stephen Soderland, Oren Etzioni[13]

This paper present G-FLOW, a multidocument summarization system aimed at generating coherent summaries. While previous MDS systems have focused primarily on salience and coverage but not coherence, G-FLOW generates an ordered summary by jointly optimizing coherence and salience. G-FLOW estimates coherence by using an approximate discourse graph, where each node is a sentence from the input documents and each edge represents a discourse relationship between two sentences. Since this is a graph based approach, relation between sentences can represent. Disadvantage is the lack of concept of abstraction.

viii) Using Sentence-based Topic Models - Dingding Wang Shenghuo Zhu Tao Li Yihong Gong[14]

This paper, propose a new Bayesian sentence-based topic model for multi-document summarization by making use of both the term-document and term sentence associations. This proposal explicitly models the probability distributions of selecting sentences given topics and provides a principled way for the summarization task. An efficient variational Bayesian algorithm is derived for estimating model parameters. It is a cluster based approach and existing algorithms are used, eg. K –means, Cosine similarity etc. Hence no need of complex procedures. But lack of semantic approach is a disadvantage.

ix) Graph-Based Iterative Ranking Algorithms and Information Theoretical Distortion Measures - Borhan Samei , Marzieh Eshtiagh[15]

This paper propose an extraction-based model which constructs a directed weighted graph from the original text by adding a vertex for each sentence, and compute a weighted edge between sentences which is based on distortion measures. Finally, a ranking algorithm is applied to identify the most important sentences to be included in the summary. Its also a simple graph based approach. Simple feature like frequency is used but abstraction is not used.

x) Summarization Based on Cluster - Khanapure V.M, Prof. Chirchi V.R[16]

In this paper it search and rank the existing cases according to users' requests in a semantic way and provide a better result representation by grouping and summarizing the retrieved past cases to make the system fully functional and usable. The high performance of multidoument summarization based on cluster using sentence-level semantic analysis (SLSS), mixture model and symmetric non-negative matrix factorization (SNMF). Its a cCluster based approach. Advantage is that semantic calculations are done to some extent. Drawback is the missing of abstraction.

IV. Conclusion

This study gives a survey on various approaches for information extraction from multiple documents. This also helps existing researchers and new ones to get an idea of information extraction from multiple documents and a comparison with advantages and disadvantages. Here we have discussed traditional approaches and compared number of existing techniques. Each of these techniques possesses its own advantages and limitations towards semantic text extraction. For future improvement, we propose a novel approach, which treat text semantically than textual form and syntax to generate a better summary which is well suited for an informative type summary generation. We believe that an ontology based approach can eliminate many limitations we have discussed so far.

References

[1] Chenghua Dang, Xinjun Luo "WordNet-based Document Summarization", in 7th wseas int. conf. on applied computer & applied computational science (acacos '08), hangzhou, china, april 6-8, 2008

[2] Md. Majharul Haque, Suraiya Pervin, and Zerina Begum "Automatic Multiple Documents Text Summarization." International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 3 No. 1 May 2013, pp. 121-129

[3] Yogan Jaya Kumar, Naomie Salim, "Automatic Multi Document Summarization Approaches", in Journal of Computer Science 8 (1): 133-140, 2012

[4] Jade Goldstein, Vibhu Mittal, Mark Kantrowitz and Jaime Carbonell, "Multi-Document Summarization by Sentence Extraction," ANLP/NAACL Workshops. Association for Computational Linguistics, Morristown, New Jersey, pp. 40-48, 2000.

DOI : 10.18231/2454-9150.2018.0607

[5] G. Erkan and D. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," Journal of Artificial Intelligence Research, vol. 22, pp. 457-479, 2004.

[6] Tiedan Zhu and Xinxin Zhao, "An Improved Approach to Sentence Ordering For Multi-document Summarization," IACSIT Hong Kong Conferences, IACSIT Press, Singapore, vol. 25, pp. 29-33, 2012.

[7] Chin-Yew Lin Eduard Hovy, "From Single to Multi-document Summarization: A Prototype System and its Evaluation" Proceedings of the ACL conference, pp. 457–464. Philadelphia, PA. 2002

[8] He Liu, Hongliang Yu, Zhi-Hong Deng "Multi-Document Summarization Based on Two-Level Sparse Representation Model" Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence,2015

[9] Wen-tau Yih Joshua Goodman Lucy Vanderwende Hisami Suzuki "Multi-Document Summarization by Maximizing Informative Content-Words", IJCAI'07 Proceedings of the 20th international joint conference on Artificial intelligence Pages 1776-1782 ,2007

[10] Xiaojun Wan, Jianwu Yang and Jianguo Xiao "Manifold-Ranking Based Topic-Focused Multi-Document Summarization" IJCAI'07 Proceedings of the 20th international joint conference on Artifical intelligence Pages 2903-2908 ,2007

[11] Jade Goldstein, Vibhu Mittal t Jaime Carbonell, Mark Kantrowitzt "Multi-Document Summarization By Sentence Extraction" , Proceeding NAACL-ANLP-AutoSum '00 Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization - Volume 4 Pages 40-48

[12] John M. Conroy, Judith D. Schlesinger and Dianne P. O'Leary "Topic-Focused Multi-document Summarization Using an Approximate Oracle Score", International Journal of Scientific Engineering and Technology Research Volume.03, IssueNo.01, January-2014, Pages:0001-0006, 2014

[13] Janara Christensen, Mausam, Stephen Soderland, Oren Etzioni "Towards Coherent Multi-Document Summarization", NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference, 2013

[14] Dingding Wang Shenghuo Zhu Tao Li Yihong Gong "Multi-Document Summarization using Sentence-based Topic Models", ACLShort '09 Proceedings of the ACL-IJCNLP 2009 Conference Short Papers Pages 297-300, 2009.

[15] Borhan Samei and Marzieh Eshtiagh "Multi-Document Summarization Using Graph-Based Iterative Ranking Algorithms and Information Theoretical Distortion Measures", Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference, 2014

[16] Khanapure V.M, Prof. Chirchi V.R "Multi-document Summarization Based on Cluster", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering Vol. 3, Issue 4, 2014.

[17] Judith D. Schlesinger, Dianne P. O'Leary and John M. Conroy, "Arabic/English Multi-document Summarization with CLASSY - The Past and the Future," Proceedings of the 9th international conference on Computational linguistics and intelligent text processing, Haifa, Israel, pp. 568–581, 2008.M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

DOI : 10.18231/2454-9150.2018.0607