CLUSTER BASED AND GRAPH BASED METHODS OF SUMMARIZATION: SURVEY AND APPROACH

Sonam Hinge¹, Sheetal Sonawane²
¹Department of Computer Engineering, PICT, Savitribai Phule Pune University Pune, India
²Department of Computer Engineering, PICT, Savitribai Phule Pune University Pune, India

ABSTRACT:

In today's world, online transfer of information is increasing rapidly. Accessing such large information on particular topic is hectic task. So, automatic summarization is the solution where summary of documents is generated. Nowadays, document summarization is the emerging technique and a lot of research is done and going on in this field. Document summarization can either be on single document or multiple documents. The common approach is analysis of word using term frequency and inverse term frequency. Cluster based document summarization and graph based document summarization are two main approach of summarization. This paper represents systematic survey on document summarization. Different graph based approach, query based approach, and cluster based approaches discussed here. The result of different approach and their analysis is presented in this paper.

Keywords: Text summarization; information retrieval; cluster based; graph based; multidocument summarization.

[1] INTRODUCTION

With rapid growth in technology, industries and so in other sectors such as living style and fashion world, online information increased rapidly. A lot of information is created and transferred daily. In information transfer, electronic media is leading to tackle constraint of time. It is necessary to transfer information effectively and efficiently. The media like news web-sites, Facebook, Twitter are bombarded with a lot of posts and news in every minute and day. In case of news articles, same event or news is handled by different news channels which give different description of event as well as information that is unique to that channel. Different channels take different angles to cover the news. The vast amount of overloaded and redundant information leads to document summarization. Automatic summarization generates a simple and precise summary easy to understand and describe overall documents. The survey covers-

- Various graph and cluster based approach for text summarization.
- Comparative study of different methods, their advantages and limitations.

The organization of this paper is as follows. Section 2 describes different types of document summarization. Section 3 describes comparative analysis of graph based and cluster based approach. Section 4 present conclusions.
[2] DOCUMENT SUMMARIZATION

Document summarization is a process of extracting information i.e. sentences from the document those are very important, preserving overall meaning of document.

Document summarizer system take input as a single document or multiple documents. It find out TF-IDF of each word and then find out sentence score using different methods such as PageRank [2, 4, 9], QoR[5], SentenceRank[3], affinity score[6]. Clustering approach is used to cluster documents [4, 9]. Graph is collection of nodes and edges where nodes are represented by using word, sentence, document etc. and relation between them is represented by using edge. Edge is connected by taking different parameters into consideration such as consecutively appearance, position, semantics of node etc. Graph approach is used to represent text document as a graph and then extract the most important document based score and ranking of sentences [1, 3, 5, 6, 8].

The main goal of a summary is to present document or set of documents is a short and readable paragraph. Summaries can be produced from single document or from many documents. Figure 1 shows different types of document summarization. These types are discussed below.

![Figure 1. Types of Document Summarization.](image_url)

The task of producing summary from document is called multidocument summarization [4, 5, 8]. Summaries are produce in two ways i.e. either extractive or abstractive. In extractive summarization approach, most important sentences are identified and generate summary without changing actual sentence. Contrast to extractive summarization, abstractive summarization works on concept of a document/text and generate new sentences from the concept. Sentence fusion, semantic analysis is mainly used to generate new sentences from original sentences. Keyword summarization is used to cover overall document into minimum keywords such as title of chapter, article, topic which reduce time to identify whether document is relevant or not.

Summarization can also be specific to the need of user, thus is called query specific summarization. For QSS, relevant documents are extracted, and then by using association between query, document, sentences, words response is given back [5]. Opinion summarization is the application of text summarization which is widely used in app specific domain to improve the application and popularity.

[3] APPROACH OF DOCUMENT SUMMARIZATION
There is a lot of research done in the area of text summarization. Many researchers worked on single document summarization [1, 2, 9] whereas many worked on multidocument summarization [4, 5, 8]. Important sentences are used to generate summary. There are different methods to calculate sentence score such as pagerank algorithm [2, 9], modified pagerank algorithm [4], Hopfield network algorithm [2], shortest path approach [8], QoR model [5].

[3.1] CLUSTER BASED APPROACH

Document clustering is very important for gathering similar documents together.

Zhang, L. Sun, and Q. Zhou [4] (2005) proposed a cue-based hub-authority approach for multi-document text summarization. In this approach, they combined the text content with some cues such as cue phrase, sentence length and first sentence and explore sub document in the multi-documents by identifying these sub-topics.

First sentence is the best predictive feature because in most of the articles, first sentence often describes more information. Cue phrase like "conclusion" or "in particular" are followed by important sentence. So, sentence containing these words or cues are more important than other sentences. Sentence length is another important feature of document. Generally, long sentences contain some more information than short sentences.

1. Firstly detect the subtopics by sentence clustering and extract feature words of different subtopics. Use KNN clustering for generate partition and TF*IDF to extract feature words.
2. Secondly, all feature words and cue phrases are used as a vertex of hub. Cue phrases are used for vertex of hub and sentences are for vertex of authority. If sentence contains the words in Hub, there is a directed edge pointing from Hub word to authority sentence.
3. Hub authority algorithm is used to rank importance of sentences within multi-document.
4. Thirdly, use Marko Model to order subtopics that final summarization should contain and generate summary according to sentence ranking.

<table>
<thead>
<tr>
<th></th>
<th>ROUGH-1</th>
<th>ROUGH-2</th>
<th>ROUGH-LCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Policy</td>
<td>0.3149</td>
<td>0.0573</td>
<td>0.1059</td>
</tr>
<tr>
<td>Position Policy</td>
<td>0.3528</td>
<td>0.08550</td>
<td>0.1293</td>
</tr>
<tr>
<td>Our approach</td>
<td>0.3761</td>
<td>0.0908</td>
<td>0.1326</td>
</tr>
</tbody>
</table>

Dataset DUC 2004 is used to test proposed system. Table 1 shows comparative improvement in result. Three of the ROUGE metrics in our experiment results: ROUGE-1 (unigram-based), ROUGE-2(bi-gram based), and ROUGE-LCS (based on longest common subsequence weighted by the length) [4].
S. S. Ge, Z. Zhang, and H. He [9] (2011) proposed weighted graph model based sentence clustering and ranking for document summarization. In proposed system, an algorithm that combines sentence ranking and clustering has been proposed for document summarization. Weighted undirected graph is used to take sentence similarities and the discourse relationship between sentences as the weights of edges is constructed for the given document. A graph ranking algorithm is used to calculate sentence score. The contribution is mainly:

1. An extract based document summarization algorithm is introduced which combines sentence ranking and clustering;
2. A sentence clustering algorithm base on SNMF is presented; and
3. A weighted graph model considering discourse relationship between sentences is presented to cluster and rank sentences in a document.

The algorithm ranks all the sentences of a document, and then clusters the sentences into different subtopic groups. High ranking sentences representing each subtopic are selected to compose the summarization. The PageRank algorithm has been used for sentence ranking, and an algorithm based on SNMF has been introduced to cluster sentences. Cosine similarity is used to measure the content overlap between sentences.

\[
W(S_i, S_j) = \frac{S_i \cdot S_j}{|S_i||S_j|}
\]

Eq.(1)

All sentences in the document are ranked by applying pagerank algorithm based on weighted undirected graph model. The rank of each vertex of the graph is calculated as follows:

\[
r(v_i) = d \sum_{j=1}^{n} r(v_j)w_{ji} + (1 - d)
\]

Eq.(2)

<table>
<thead>
<tr>
<th>Methods</th>
<th>ROUGH-1</th>
<th>ROUGH-2</th>
<th>ROUGH-L</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>0.30955</td>
<td>0.05411</td>
<td>0.27792</td>
</tr>
<tr>
<td>SR+KC</td>
<td>0.30384</td>
<td>0.05233</td>
<td>0.27114</td>
</tr>
<tr>
<td>SR+NMF</td>
<td>0.31170</td>
<td>0.05660</td>
<td>0.27890</td>
</tr>
<tr>
<td>SR+SNMF</td>
<td>0.32042</td>
<td>0.06440</td>
<td>0.28520</td>
</tr>
</tbody>
</table>

To investigate the effect of different clustering algorithm, the performances of different component combination of proposed algorithm are compared in Table. Table 2 shows analysis of performance of sentence rank algorithm with combination of different clustering algorithm [9].

### [3.2] QUERY BASED APPROACH

Query based approach is used where user asked information about particular topic and the response is depends on number of documents. So, generate summary which is close to the user query. This type of approach is mainly used in search engine.
Furu Weu, Yanxiang He, Wenjie Li, Qin Lu [2] (2008) proposed A Query Sensitive Graph-Based Sentence Ranking Algorithm for Query-Oriented Multi Document Summarization. The main contribution of this paper is:
1. To introduce the query-sensitive similarity measure into the existing graph model for sentence edge weight estimation;
2. Differentiate the intra-document and inter document sentence relations and consider the influence of entire document context on individual sentence evaluation.

Table 3. Performance evolution of QssR[2]

<table>
<thead>
<tr>
<th></th>
<th>ROUGH-1</th>
<th>ROUGH-2</th>
<th>ROUGH-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>QssR</td>
<td>0.37486</td>
<td>0.07525</td>
<td>0.13272</td>
</tr>
<tr>
<td>S15</td>
<td>0.37383</td>
<td>0.07251</td>
<td>0.13163</td>
</tr>
<tr>
<td>S17</td>
<td>0.36901</td>
<td>0.07174</td>
<td>0.12972</td>
</tr>
<tr>
<td>S10</td>
<td>0.36640</td>
<td>0.07089</td>
<td>0.12649</td>
</tr>
<tr>
<td>NIST BASELINE</td>
<td>0.30217</td>
<td>0.04947</td>
<td>0.09788</td>
</tr>
</tbody>
</table>

Table 3 provides the comparison of our model with the DUC 2005 participating systems, where S15, S17 and S10 are the top three best performing systems. It clearly shows that the proposed model outperforms the first-ranked system in DUC 2005 with 3.78% increase in ROUGE-2. The results are definitely promising, since the best system (S15) is only 1.07% above the second system (S17) on ROUGE-2.

[3.3] GRAPH BASED APPROACH

Graph is a collection of vertices and edges. Graphs are important to provide structural and semantic relation between documents, sentences, words etc.[12].

Ohm Sornil, Kornnika Gree-ut [5] (2006) proposed an automatic text summarization approach using content based and graph-based characteristics. In this system, single document summarization combines content based and graph based approach and introduces the Hopfield network algorithm as a technique for ranking text segment. Here, mainly two stages considered.
1. Segments are represented by content based feature vectors.
2. Segments are represented as nodes and relationship between two segments whose similarity score above threshold are represented as edge.

Finally, graph search technique is used to recommend segment to be extracted as summary.

Table 4. Performance of PageRank and Hopfield network with different comparative rate [5]

<table>
<thead>
<tr>
<th></th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hopfield</td>
<td>0.4643</td>
<td>0.50086</td>
<td>0.52719</td>
</tr>
<tr>
<td>PageRank</td>
<td>0.42366</td>
<td>0.46083</td>
<td>0.48716</td>
</tr>
</tbody>
</table>

Table 4 shows performance of PageRank and Hopfield algorithm with different comparison rate. The result shows Hopfield algorithm work better than PageRank.
K. S. Thakkar, R. V. Dharaskar, and M. Chandak [8] (2010) proposed Graph-based algorithms for text summarization. When a text is to be summarized, it is first split into sentences and words. The sentences become the nodes of the graph. Sentences that are similar to each other have an edge between them. Here, similarity simply means word overlap, though other measures could also be used. Thus, if two sentences have at least one word in common, there will be an edge between them. The cost of an edge from the node representing sentence number $i$ in the text, $S_i$, to the node for $S_j$ is calculated as:

$$Cost(i, j) = \frac{(i - j)^2}{\text{overlap}(i, j) \cdot \text{weight}(j)}$$  \hspace{1cm} \text{Eq}(3)$$

When the graph has been constructed, the summary is created by taking the shortest path that starts with the first sentence of the original text and ends with the 1st sentence this will hopefully give a smooth but shorter set of sentences between these two points.

Shortest path algorithm is easy to implement and is not require any language knowledge or domain knowledge. The proposed algorithm works well as compare with existing TextRank algorithm.

Niladri Chatterjee, Amol Mittal and Shubham Goyal [1] (2012) proposed Single Document Extractive Text Summarization using Genetic Algorithms. A document is represented as a weighted directed acyclic graph where each sentence is represented by a vertex and edge is occurs between two sentences if one sentence is preceded by another sentence.

A fitness function is used to express quality of a summary such as topic relation, cohesion and readability. GA is used to increase fitness function and extract important sentences. The proposed algorithm can provide query and non-query based summary.

Yang Wei [6] (2012) proposed document summarization method based on heterogeneous graph. The method is first implemented by constructing a graph which reflect relationship between different size of granularity nodes, and then using ranking algorithm to calculate score of nodes. The granularity includes words, sentences and topics. Finally, highest score of sentences in the document will be chosen as summary.

Animesh Ramesh [3] (2014) proposed SentenceRank - A Graph based approach to summarize text. He introduces a graph and an intersection based technique which uses statistical and semantic analysis for computing relative importance of textual units in large data sets in order to summarize text[3]. Current implementations consider only the mathematical/statistical approach to summarize text. (like frequency, TFIDF, etc.) But there are many cases where two completely different textual units might be semantically related. This problem is overcome by exploiting the resources of WordNet and by the use of semantic graphs dissimilarity between any pair of sentences.

Table 5. Comparative study of different algorithm using precision and recall
Table 5 shows comparison precision and recall of SentenceRank algorithm with other algorithms.

**[4] COMPARATIVE STUDY**

Different methods are used to generate summary. The summary produced by these papers is based on single document or multiple documents. Table 6 and Table 7 show comparative study of different papers.

**Table 6. Comparative study of cluster based and query based approach of document summarization**

<table>
<thead>
<tr>
<th></th>
<th>Cluster based</th>
<th>Query Based</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Paper</strong></td>
<td><a href="2005">4</a></td>
<td><a href="2011">9</a></td>
</tr>
<tr>
<td><strong>Vertex</strong></td>
<td>Feature word cue phrase</td>
<td>Sentence</td>
</tr>
<tr>
<td><strong>Edge weight</strong></td>
<td>Edge between hub and authority. Weight of authority is considered</td>
<td>Cosine similarity and discourse relationship</td>
</tr>
<tr>
<td><strong>Sentence Score</strong></td>
<td>Modified pagerank i.e. hub authority framework</td>
<td>Pagerank</td>
</tr>
<tr>
<td><strong>Graph Model</strong></td>
<td>Directed weighted graph</td>
<td>Undirected weighted graph</td>
</tr>
<tr>
<td><strong>Theme</strong></td>
<td>It uses text content with some cues e.g. cue phrases, sentence length, first sentence etc.</td>
<td>It uses sentence ranking and clustering method extract essential sentences from document</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>It combine both surface feature and content feature</td>
<td>Performance is improved due to hybrid approach</td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
<td>Difficult to detect subtopic in a document</td>
<td>Influence to system performance by weight balance &amp; sparseness</td>
</tr>
</tbody>
</table>
Table 6. Comparative study of graph based approach of document summarization

<table>
<thead>
<tr>
<th>Paper</th>
<th>Graph based approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Paper</td>
</tr>
<tr>
<td></td>
<td><a href="2006">5</a></td>
</tr>
<tr>
<td></td>
<td><a href="2010">8</a></td>
</tr>
<tr>
<td></td>
<td><a href="2012">1</a></td>
</tr>
<tr>
<td></td>
<td><a href="2012">6</a></td>
</tr>
<tr>
<td></td>
<td><a href="2014">3</a></td>
</tr>
</tbody>
</table>

[5] APPLICATIONS

Nowadays, document summarization is emerging technique due to overwhelming information. Document summarization is the technique of information retrieval and it has many applications.

1. Search engine: Google, yahoo etc. search engine provide systematical response of user’s query.
2. News channel: News about one event may take place from different angle by different reporters. To find out most important part and uniqueness to cover overall story about that event, document summarization is very important.
3. Video summarization and Image summarization: In image summarization the system finds the most similar and important images while in video summarization, remove the boring or repetitive scenes. Extract shorter and concise scenes from the video.

4. Spoken document summarization: To efficiently reduce the redundant contents and summarizes concise spoken sentences containing important words with semantic dependency.

[6] CHALLENGES

Document summarization is very important in the domain of text processing. There are number of document those are related to each other or different from each other. From pre-processing to summary generation, there are different challenges:

1. Develop efficient algorithm to summarize multiple documents.
2. Same document or different document contains redundant sentences.
3. Identify redundancy in the documents.
4. Need of pre-processing.
5. While applying semantic analysis, proper semantic should get selected otherwise it may change meaning of that sentence.
6. In case of abstractive summarization, meaning of sentences should not altered by replacing these sentences with new sentence.
7. To generate summary find out important sentences and selecting some of them to include in summary, as per threshold.
8. Generated summary should not change the meaning of documents.

[7] CONCLUSION

Document summarization has widely demand due to increase in technology and social media. It is evolving field in the domain of data mining and information retrieval and natural language processing. There are various challenges while doing processing on document, extracting important sentences from document, finding relevance in between documents etc. Different approach based on graph, query and cluster are discussed in this paper. Term frequency and inverse document frequency is the most important part in finding importance of word. Cluster based document summarization based on grouping similar document into one group. The accuracy of cluster is depends on intra- cluster similarity and inter cluster similarity. Graph based document summarization approach is important to gain richer document summary.

REFERENCES


