DATA SUMMARIZATION USING MAXIMAL MARGINAL RELEVANCE METHOD

Sandeep Kumar Mathariya, Vishakha Soni, Anand Sen and Ranu Soni
Patel College of Science and Technology, Rala Mandal Indore
E-mail: mathariya@gmail.com; senanand05@gmail.com; soni.vishakha@gmail.com

Abstract: The search for interesting information in a huge data collection is a tough job frustrating the seekers for that information. The automatic text summarization has come to facilitate such searching process. Automatic text summarization is to compress an original document into an abridged version by extracting almost all of the essential concepts with text mining techniques. The selection of distinct ideas “diversity” from the original document can produce an appropriate summary. Incorporating of multiple means can help to find the diversity in the text.

Here, the approach for text summarization, in which three evidences are employed (clustering, binary tree and diversity based method) to help in finding the documents distinct ideas. The emphasis of the approach is on controlling the redundancy in the summarized text. The role of clustering is very important, therefore used K-means clustering algorithm. Document-Sentence Tree is build on which the MMI approach is applied to get the quality text Summary.

1. INTRODUCTION

Due to the rapid growth of the World Wide Web, information is much easier to disseminate and acquire than before. Finding useful and favored documents from the huge text repository creates significant challenges for users. Typical approaches to resolve such a problem are to employ information retrieval techniques. Information retrieval relies on the use of keywords to search for the desired information. Nevertheless, the amount of information obtained via information retrieval is still far greater than that a user can handle and manage. This in turn requires the user to analyse the searched results one by one until satisfied information is acquired, which is time-consuming and inefficient. It is therefore essential to develop tools to efficiently assist users in identifying desired documents. One possible means is to utilize automatic text summarization. Automatic text summarization is a text-mining task that extracts essential sentences to cover almost all the concepts of a document. It is to reduce users’ consuming time in document reading without losing the general issues for users’ comprehension. With document summary available, users can easily decide its relevancy to their interests and acquire desired documents with much less mental loads involved.

In automatic text summarization, the selection process of the distinct ideas included in the document is called diversity. The diversity is very important evidence serving to control the redundancy in the summarized text and produce more appropriate summary. Many approaches have been proposed for text summarization based on the diversity. The pioneer work for diversity based text summarization is MMR (maximal marginal relevance), it was introduced by Carbonell and Goldstein [2], MMR maximizes marginal relevance in retrieval and summarization. The sentence with high maximal relevance means it is highly relevant to the query and less similar to the already selected sentences. Our modified version of MMR maximizes the marginal importance and minimizes the relevance. This approach treats sentence with high maximal importance as one that has high importance in the document. This approach employs two evidences (clustering algorithm and a binary tree) to exploit the diversity among the document sentences. A procedure for creating approximate structure for document sentences in the form of a binary tree, in our study, we build a binary tree for each cluster of document sentences, where the document sentences are clustered using a clustering algorithm into a number of clusters equal to the summary length. An objective of using the binary tree for diversity analysis is to optimize and minimize the text representation; this is achieved by selecting the most representative sentence of each sentences cluster. The redundant sentences are
prevented from getting the chance to be candidate sentences for inclusion in the summary, serving as penalty for the most similar sentences. MMI is used to select one sentence from the binary tree of each sentence cluster to be included in the final summary.

2. LITERATURE REVIEW

Many approaches have been proposed for text summarization based on the diversity. The pioneer work for diversity based text summarization is MMR (maximal marginal relevance), it was introduced by Carbonell and Goldstein [2]. MMR maximizes marginal relevance in retrieval and summarization. The sentence with high maximal relevance means it is highly relevant to the query and less similar to the already selected sentences.

MMR has been modified by many researchers [4; 10; 12; 13; 16; 21; 23]. Our modification for MMR formula is similar to Mori et al.'s modification [16] and Liu et al.’s modification [13] where the importance of the sentence and the sentence relevance are added to the MMR formulation. Ribeiro and Matos [19] proved that the summary generated by MMR method is closed to the human summary, motivating us to choose MMR and modify it by including some documents features. Here it employs two evidences (clustering algorithm and a binary tree) to exploit the diversity among the document sentences. Neto et al. [17] presented a procedure for creating approximate structure for document sentences in the form of a binary tree, build a binary tree for each cluster of document sentences, where the document sentences are clustered using a clustering algorithm into a number of clusters equal to the summary length. An objective of using the binary tree for diversity analysis is to optimize and minimize the text representation; this is achieved by selecting the most representative sentence of each sentences cluster. The redundant sentences are prevented from getting the chance to be candidate sentences for inclusion in the summary, serving as penalty for the most similar sentences. Employed idea is similar to Zhu et al.’s idea [25] in terms of improving the diversity where they used absorbing Markov chain walks.

3. PROBLEM STATEMENT

A method for summary generation depending on the extraction of the highest important sentences from the original text, by introducing a modified version of MMR (Maximal Marginal Relevance) called as MMI (maximal Marginal Importance) diversity based approach.

4. OBJECTIVES & SCOPE

1. To get the most close & relevant text summary from the document.
2. To get a method that uses an effective diversity based method for document summarization called as Maximal Marginal importance (MMI).
3. To use maximum sentence features to generate accurate summary than the MMR.
4. To show the importance of the sentence to emphasize on the high information richness in the sentence as well as high information novelty.
5. To employ multiple factors helping to find the diversity in the text.
6. To maximize the coverage of each sentence by taking into account the sentence relatedness to all other document sentences.

6. METHODOLOGY

The method for summary generation depends on the extraction of the highest important sentences from the original text, hence introduced a modified version of MMR, and called it MMI (maximal marginal importance). MMR approach depends on the relevance of the document to the query, and it is for query based summary. In the modification, tried to release this restriction by replacing the query relevance with sentence importance for presenting the MMI as generic summarization approach. Most features used are accumulated together to show the importance of the sentence, the reason for including the importance of the sentence in the method is to emphasize on the high information richness in the sentence as well as high information novelty. The tree for grouping the most similar sentences together in easy way are used, and assumed that the tree structure can take part in finding the diversity.

MMI is used to select one sentence from the binary tree of each sentence cluster to be included in the final summary. In the binary tree, a level penalty is imposed on each level of sentences which is 0.01 times the level number. The purpose of the level penalty is to reduce the noisy sentences score. The sentences which are in the lower levels are considered as noisy sentences because they are carrying low scores. Therefore the
level penalty in the low levels is higher while it is low in the high levels. It is assumed that this kind of scoring will allow to the sentence with high importance and high centrality to get the chance to be a summary sentence, this idea is supported by the idea of PageRank used in Google [1] where the citation (link) graph of the web page or backlinks to that page is used to determine the rank of that page. The summary sentence is selected from the binary tree by traversing all levels and applying MMI on each level sentence.

\[
MMI(S_i) = \text{Arg} \max_{S_j \in \mathcal{S}} \left( \text{Score}_{\lambda}(S_i) - \beta(S_j) - \max_{S_j \in \mathcal{S}} \left( \text{Rel}(S_i, S_j) \right) \right)
\]

Where Rel\((S_i, S_j)\) the relevance between the two competitive sentences, \(S_j\) is the unselected sentence in the current binary tree, \(S_i\) is the already selected sentence, SS is the list of already selected sentences, CS is the competitive sentences of the current binary tree and \(\beta\) is the penalty level. In MMR, the parameter \(\lambda\) is very important; it controls the similarity between already selected sentences and unselected sentences, and where setting it to incorrect value may cause creation of low quality summary. The method pays more attention for the redundancy removing by applying MMI in the binary tree structure. The binary tree is used for grouping the most similar sentences in one cluster, so we didn’t use the parameter \(\lambda\) because we just select one sentence from each binary tree and leave the other sentences.

The method is used for single document summarization as well as multi documents summarization, where it has the ability to get rid of the problem of some information stored in single document or multi-documents which inevitably overlap with each other, and can extract globally important information. In addition to that advantage of method, it maximizes the coverage of each sentence by taking into account the sentence relatedness to all other document sentences. The best sentence based on the method policy is the sentence that has higher importance in the document, higher relatedness to most document sentences and less similar to the sentences already selected as candidates for inclusion in the summary.

7. CONCLUSION

In this paper, we have presented an effective diversity based method for document summarization. Two ways are used for finding the diversity: the first one is as preliminary way where the document sentences are clustered based on the similarity - similarity threshold is 0.03 determined empirically - and all resulting clusters are presented as a tree containing a binary tree for each group of similar sentences. The second way is to apply the proposed method on each branch in the tree to select one sentence as summary sentence. The clustering algorithm and binary tree is used as helping factor with the method for finding the most distinct ideas in the text. The results of the method supports that employing of multiple factors can help to find the diversity in the text because the isolation of all similar sentences in one group can solve a part of the redundancy problem among the document sentences and the other part of that problem is solved by the diversity based method which tries to select the most diverse sentence from each group of sentences as compared to the MMR approach used earlier. The advantages of the introduced method are: it does not use external resource except the original document given to be summarized and deep natural language processing is not required. The method has shown good performance when comparing with the benchmark methods used in the study. For future work is to plan to incorporate artificial intelligence technique with the proposed method.

References


