A GRAPH BASED CONCEPTUAL MINING MODEL FOR ABSTRACTIVE TEXT SUMMARIZATION

Akshata Rahul Pathak

Dempo Higher Secondary School of Science, Pace

ABSTRACT

The main objective of automatic text summarization is to compress the document into a smaller version by preserving the important concepts. This work proposes a hybrid approach of Singular Value Decomposition and Named Entity Recognition to extract important sentences present in a document. The extracted sentences are used to create a probabilistic graphical model called a Belief network. This graph model represents document summary in concept level. We have used a modified Page Rank algorithm to find the most ranked noun phrase. From this noun phrase we extracted the most relevant sentences. Findings: Our abstractive graph based model for a document generates novel sentences as it uses the concept of triplets (Subject, Verb, and Object). It identifies whether a sentence is created by structural rearrangement of another sentence. Using SVD (Singualr Value Decomposition) and NER (Named Entity Recognition) we extracted relevant information present in a document so that entire document is crushed in to a graph model. We can use this model for documents similarity as well as for plagiarism detection. Experimental results of our proposed system show that use of named entities and SVD increases the accuracy of summarizer.

INTRODUCTION

The amount of printed data increases the need for automatic summarization. A huge amount of data is accessible on the web; however, to deal with the required data is an endless task. The aim of document summarization is to extract the consolidated adaptation of the initial document. An overview of the document is valuable because it can provide an analysis of the initial document in less time. Readers can choose whether or not to peruse the full archive in the wake of spending the outline. For instance, before reading the full paper initially readers will read the abstract of a logical article. Web indexes additionally utilizing synopses of content to help clients settle on significant choices1. There are two types of document summarization techniques: extractive and abstractive. In extractive approaches summary is created by connecting extracts taken from the corpus, whereas in abstraction novel, sentences are created by extracting information from the corpus. This work focus on an abstractive summary for a document. This work proposes a hybrid approach of Singular Value decomposition and Named Entity Recognition to extract important sentences present in a document. We have used SVD to identify various relationships in a text document. In our work the document is divided to multiple documents and SVD is applied to the term-document matrix created from these documents. From SVD we can identify the most significant paragraph present in a document. Named Entity Recognition is one of the text analysis processes in machine learning. NER is defined as the subtask of information extraction that classifies elements in text into pre specified categories such as persons, locations, organizations, etc. Using NER we identified important named entities present in a document. This improves the accuracy of documentsummary. A sentence similar to title of the document is also extracted.

PREVIOUS WORKS

In paper2, a unique approach is shown to make an abstractive summarization for a particular record using a rich semantic diagram diminishing method. The strategy summaries the data record by making a rich semantic diagram for the first archive, decreasing the produced chart, and afterward creating the abstractive rundown from the diminished diagram. Paper3, the similar sentences that belongs to the same event of news articles are grouped. The brief summary is produced using language generation. In their approach, the first comparable sentence is proposed utilizing a shallow parser and after that sentence are mapped to predicate-contention structure, the content organizer utilizes subject convergence calculation to decide basic expressions. FUF/SURGE generator utilizes by sentence generation phase to join and organizes the chose phrases into new summary. In paper4, the utilization of particular quality disintegration (SVD) in content outline is determined. Firstly, mention the scientific classification of content outline strategies. At that point interpret the standards of the SVD and its conceivable outcomes to recognize semantically vital parts of content. In the second segment proposes two new assessment approaches, taking into account SVD, which gauge content examination between a unique archive and its synopsis. In assessment way, synopsis methodology is contrasted and 5 other accessible summarizers. Finally, learn the effect of the outline length on its quality from the point of the three assessments separated from an established substance based evaluator; both recently created SVD-based evaluators. At last, contemplate the impact of the outline length on its quality from the point of the three assessments. In Paper5, a measurable methodology for content synopsis joins the K Mixture expression weighting plan, which constructs itself in light of a numerical (probabilistic) ground. The phonetic strategy that investigates term connections by finding the semantic relationship noteworthiness of things and sentence semantics. Two trials are directed to confirm the proposed approach. The output of Experiment I demonstrates that even a straightforward TF-IDF term weighting strategy can upgrade the arrangement execution while the after effects of Experiment II demonstrate that proposed approach, KSRS, performs best, which legitimizes the possibilities of KSRS in the content Summarization applications. In the extension, with KSRS, this work picks a lower outline extent without stressing over the execution disintegration. In paper6, a structure was recommended for creating an abstractive outline from a semantic model of a multimodal report. The structure has three stages. Initial step a semantic model is built utilizing information representation taking into account objects, composed by ontology. In the second step, data substance is evaluated on data density matrix. In the third step, the vital ideas are imparted as sentences. The expressions served by the parser are secured in a semantic model for communicating ideas and relationship. A vital purpose of enthusiasm of this structure is that it produces theoretical synopsis, whose scope is impressive in light of the printed and graphical substance from the whole archive. In paper7 domain ontology of news events are explained by domain experts. News corpus and Chinese news words are used to create significant terms for the document pre-processing stage. Based on fuzzy ontology, news summarization is created by news agents. The advantage of this methodology is that it misuses fuzzy ontology to handle unverifiable data that simple domain ontology cannot. Few restrictions are there for this approach. This methodology might not be appropriate for English News because it is restricted only to Chinese news. And domain ontology, Chinese word reference and news corpus must be characterized by a field expert which is time expending.

SUMMARIZATION METHOD

Text summarization is the procedure of compressing an initial document into a smallest form by extracting the most relevant data out of the document. Text summarization helps to identify the most critical sections or sentences from a given record, by eliminating unimportant information, instead of keyword extraction. The Figure 1 represents the system architecture.

Figure 1. System architecture.

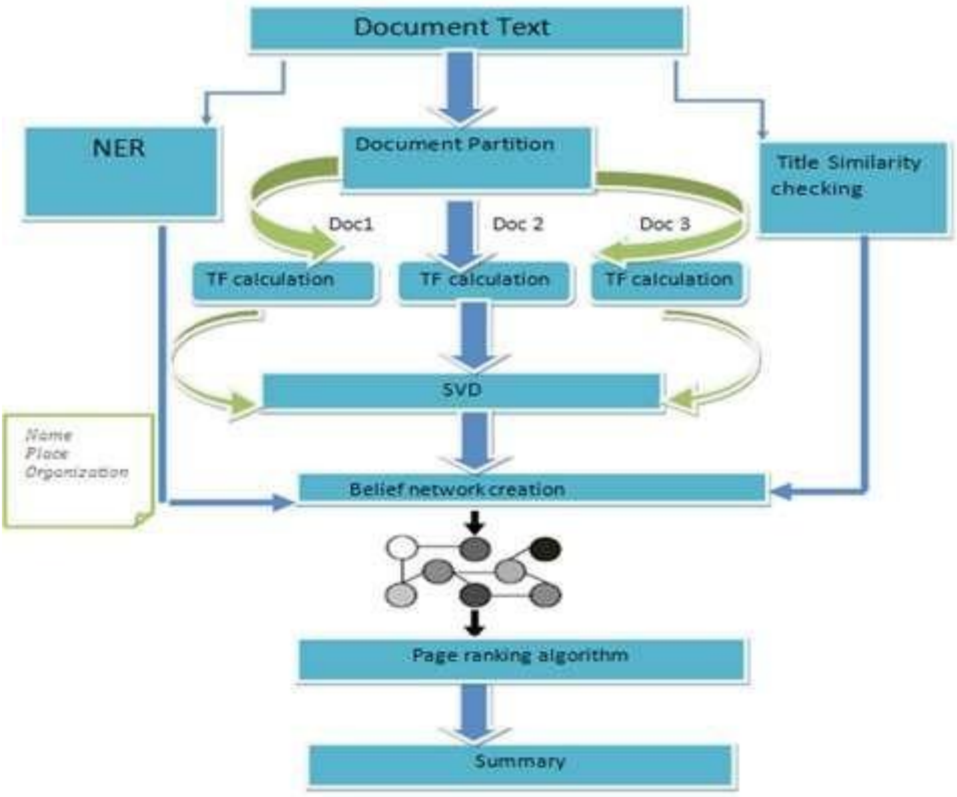


Figure 1. System architecture.

Pre processing

Pre-processing is an important module in the summarization. This pre-processing step includes tokenization, stop word removal, paragraph division.

Tokenize

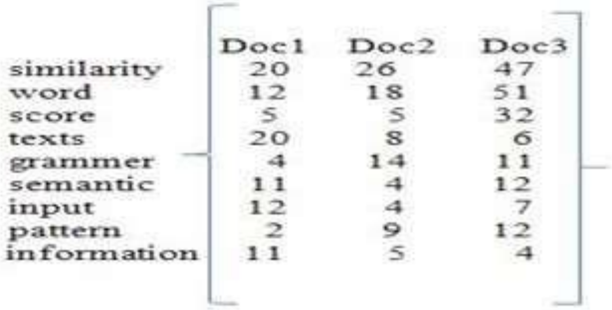
Tokenization is the procedure of partitioning a document into words, expressions, images, or other significant components called tokens. The point of the tokenization is the investigation of the words in a sentence. The rundown of tokens gets to be contributing for further handling, for example, parsing or text mining. The fundamental utilization of tokenization is recognizing the significant catch phrases.

Stop Word Removal

Stop words are regularly utilized basic words like “and”,” are”,” this” and so on. They are not valuable in groups of records. So they should be evacuated. Be that as it may, the advancement of such stop word rundown is troublesome and conflicting between literary sources. This procedure likewise diminishes the content, information and enhances the framework execution. Every content archive manages these words which are a bit much for content mining applications.

Partition

The input text is divided into different paragraphs and stored as different documents by considering the total word count of the text document. Apply Term Frequency into the multi documents. Term frequency, weight is a statistical measure used to evaluate how important a word in a document.



	Doc1	Doc2	Doc3
similarity	20	26	47
word	12	18	51
score	5	5	32
texts	20	8	6
grammer	4	14	11
semantic	11	4	12
input	12	4	7
pattern	2	9	12
information	11	5	4

Figure 2. Term frequency matrix for SVD calculation.

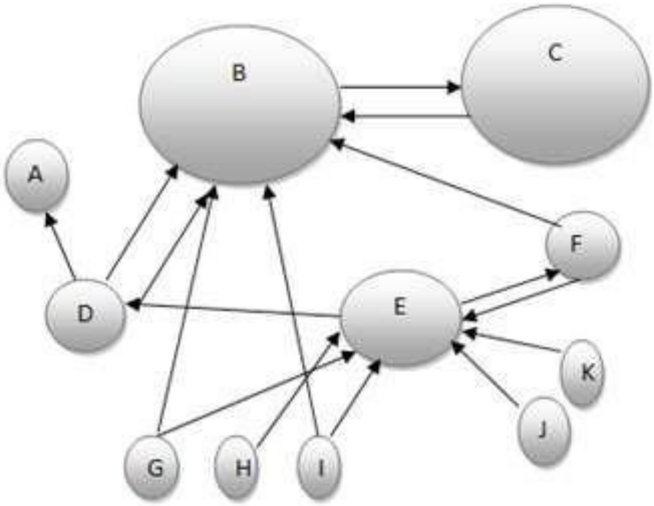


Figure 3. Page ranking representation.

Figure 2 shows the words and their corresponding term frequency, weight is indicated in the matrix form. Term frequency, weight matrix is the input for Singular Value Decomposition (SVD).

SVD Based Summarization

Singular Value Decomposition (SVD) to recognize patterns in the connections between the terms and ideas contained in an unstructured accumulation of content. SVD is a strategy that models connections between words and sentences. It has the capacity of noise reduction, which prompts a change in precision. A generalized diagonalization procedure that will allow us to “diagonal matrix – square or not square, invertible or not invertible is called the s value decomposition.

The definition of singular value decomposition. Let A be the term-document matrix of order mxn computed using term frequency.

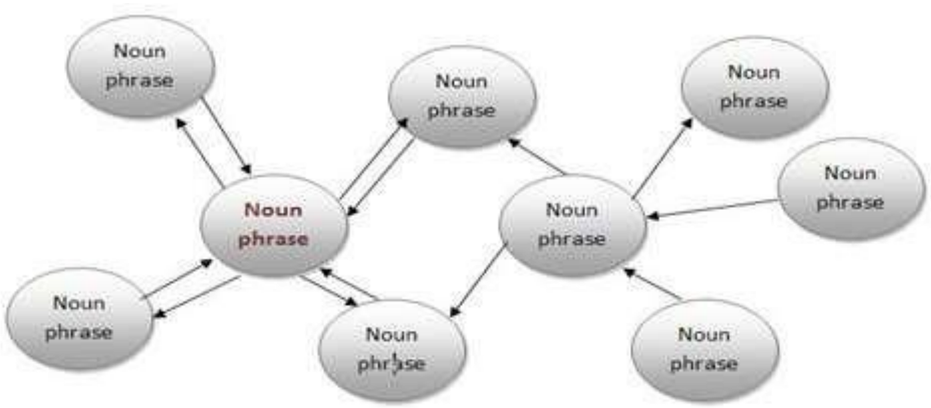


Figure 4. Ranking of Noun phrase.

A[n x m] = U[n x r]S [r x r] (V[m x r])T

Where U is a m × m orthogonal matrix, V is an m × r orthogonal matrix. Singular value decomposition is a mathematical method which models the relationships among terms and sentences. It decomposes the input matrix into The SVD basically identifies concepts present in a matrix. Calculating the SVD consists of finding the eigenvalues and eigenvectors of AAT and ATA. The columns of V are the eigenvectors of ATA and eigenvectors of AAT make up the columns of U. Matrix U represent importance of concept in a document and V represent concept term relationship. From U we can identify the most important paragraph present in a document and V provide the important term in a concept.

NER Based Summarization

Named Entity Recognition based summarization is the third module in this proposed approach. Named Entity Recognition is one of the text analysis processes in Text Mining. NER module is used to identify named entities in text into pre-defined categories such as the names of persons, organizations and locations. A news article often reports an event that can be effectively summarized by who, when and where approach. Many of these are associated with appropriate named entities in the article. It is important to include the names of algorithms used in the document summary of research article.

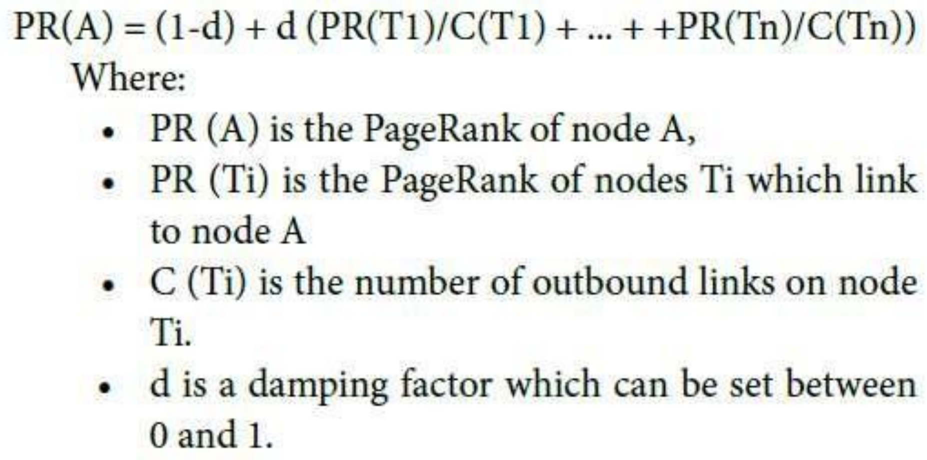
Similarity Checking

Title similarity checking is conducted in the proposed approach. The title of the document is compared to the sentences if similar sentences can be found; it is also included in the document summary.

Belief Network Creation

A belief network15 is a probabilistic graphical model, which uses a directed acyclic graph (DAG). DAG demonstrates arrangement random variables and their conditional dependencies. Here a sentence is decomposed to triplets (subject, verb, object).In our belief network nodes are either subject or object. The conditional dependency, explains the connections between the subjects and objects. After applying SVD and NER the extracted sentences are used to create the Belief network. This network represents the summary of the original document. Page Rank Algorithm is applied to the created belief network to rank nodes present in the network. To rank websites in their search engine results Google Search uses

Algorithm 1 Page Rank Algorithm[8]
Input: Generated graph
Output:Relevant noun phrases



$$PR(A) = (1\text{-}d) + d\,(PR(T1)/C(T1) + \ldots + +PR(Tn)/C(Tn))$$

Where:

- PR (A) is the PageRank of node A,
- PR (Ti) is the PageRank of nodes Ti which link to node A
- C (Ti) is the number of outbound links on node Ti.
- d is a damping factor which can be set between 0 and 1.

Page Rank algorithm. Page Rank works by counting the number and quality of links to a node to determine a rough estimate of how important the node is. The underlying assumption is that more important nodes are likely to receive more links from other nodes. PageRanks9 for a simple network is shown in figure 4 expressed as percentages. The one link to C comes from an important page B Page C has a higher Page Rank than Page E.

The Figure3 which represents the noun phrases is extracted from the tree. Most ranked noun phrases are identified by applying the page ranking algorithm. The most ranked noun phrases are calculated by counting the inbound and outbound links. The identified noun phrase sets are used to extract most appropriate sentences to form summary.

EXPERIMENTS AND EVALUATION

The experimental setup we collected 100 articles from https://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives. These documents are the input for the experiment. We have used Python language for the proposed system. For singular value decomposition we are using numpy, matrx, mat-library. This provides an expected SVD result. It is helpful for the identification of accurate keywords. SVD provides more accuracy to the selection of keywords. Named Entity Recognition can be done using a pattern library of python. NER identifies person, place, headings that includes in the summary generation. Title similarity can be checked on the document. The sentences that are similar to the title are selected and belief network be created. The SVD identifies the accurate keyword present in the document and we extracted the sentences which contains the keyword. We then created the triplets (subject, object, verb) from each sentence which is represented in the figure 5. The Page ranking algorithm is applied to rank the sentences and on the basis of that score sentences be selected and adds to the summary. Summary assessment is subjective in nature. It is exceptionally hard to figure out if a synopsis is great. For assessing synopses, both human assessment strategies and programmed (machine-based) assessment techniques are utilized. A human evaluation is done by differentiating system delivered blueprints and reference/model summaries by human judges. As shown by some predefined rules, the judges select a score in a predefined scale to each once-over under appraisal. Quantitative scores are given to the summaries considering the assorted subjective components, for instance, information substance, commonality, and so on. The real issues with human appraisal are: The assessment procedure is endless

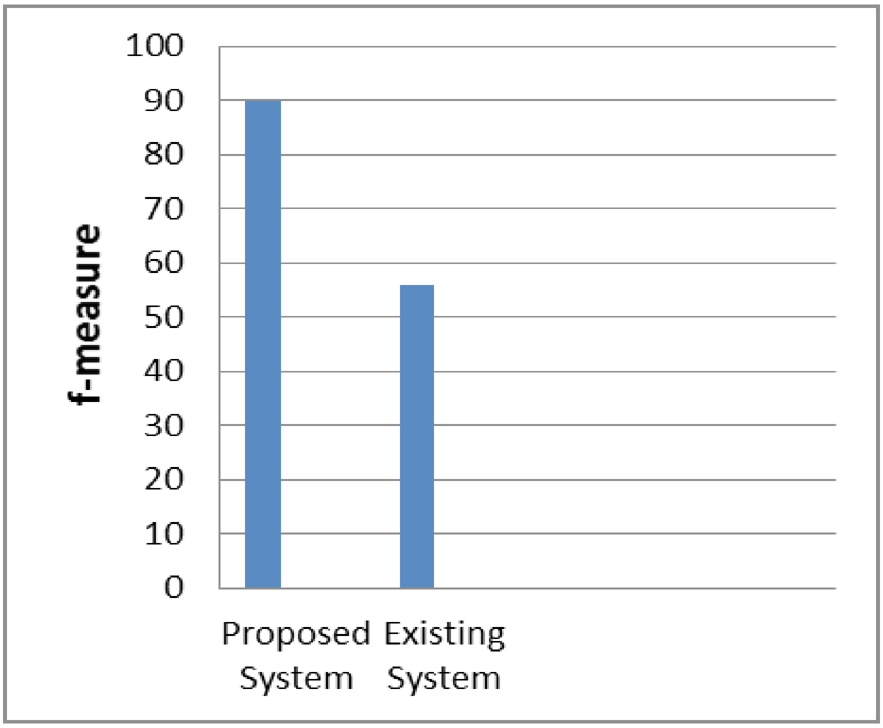
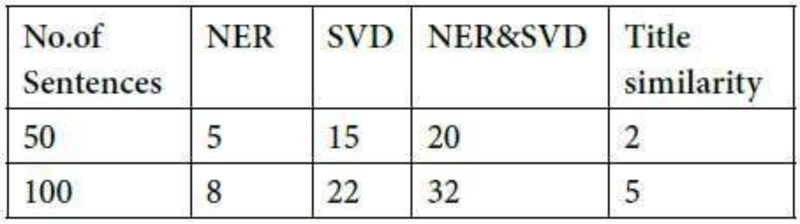


Figure 5: Efficiency measure.

We can form different sentences corresponding to the subject and object. Here Word Net can be used to replace verb. So that we can form sentences in different structures. Graph modeling can be helpful for rephrasing sentences and comparing with other online tools it is an advantage to our system.

Table 1. Sentences extracted for summary generation



No.of Sentences	NER	SVD	NER&SVD	Title similarity
50	5	15	20	2
100	8	22	32	5

The Table 1 is the visualization of sentences extracted through different methods represents the comparison of existing systems with the proposed system. Proposed system has the highest efficiency while comparing with an online tool.

CONCLUSION

We introduce an innovative summarization model, consisting of SVD-NER based summarization and Belief network creation. The Pre-processing phase followed by the SVD summarization extract the keywords accurately from the input. Then the model generated belief network using the sentences extracted from these keywords. This model implemented the page ranking algorithm to extract most relevant noun phrases. NER and Title similarity can also be proposed in the system. The future work includes Co reference Resolution, which improves the accuracy of document similarity and natural language processing.