Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 532 – 539

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Paragraph Ranking Based On Eigen Analysis

Reshma O.K.*, P.C.Reghu Raj

Dept. of Computer Science and Engineering, Govt. Engineering College, Sreekrishnapuram, Palakkad, India-678633

Abstract

The information contained in the document can be retrieved from its most significant paragraph, rather than by reading the whole document. The proposed work ranks the paragraphs of a text document using eigen analysis and returns the most important paragraph of a document. The importance of each paragraph is determined based on the correlation between the paragraphs. The proposed method explores the use of fuzzy graphs in capturing the inter-paragraph correlation of text documents. This approach models the document as a fuzzy graph where a node refers to a paragraph and an edge indicates the relationship between the paragraphs. The correlation between paragraphs is measured by extracting their semantic similarity. The importance of each node is determined based on this correlation. Subsequently the system ranks the paragraphs according to their importance. The proposed system is evaluated using DUC 2001 data set. The ROUGE scores show that the significant paragraph suggested by the proposed method covers relatively a good amount of relevant information in the document.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Paragraph correlation, Paragraph ranking, Fuzzy graph, Eigen analysis.

1. Introduction

The tremendous amount of information available has raised a matter of concern to identify the relevant information from the whole lot. The science research articles usually contain an abstract which gives a quick overview of the article, though it carries less information than full text article. However, various text summarizers are available that generate extractive summaries from the text document. Since, these summaries are extracted from different parts of a text, it may not always offer fluency and readability. Hence for an efficacious reading, brief text carrying relevant information of a full text article with fair readability is required for authors.

The essential information contained in the document can be retrieved from its most important paragraph, rather than by reading the whole document. The proposed work returns the most important paragraph of a document by ranking the paragraphs according to their importance level. This approach divides the paragraph ranking task into three sub-tasks: (1) pre-processing, (2) computing inter-paragraph correlation and (3) ranking paragraphs according to their importance.

* Corresponding author. Tel.: +91-944-641-4515.
E-mail address: okreshma@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.082

The proposed approach presents a novel paragraph ranking method to perform the final sub-task. The idea of inter-paragraph correlation is extended to model a document as a fuzzy-graph on which the paragraph ranking algorithm is proposed. The inter-paragraph correlation is captured in the second sub-task is the input to the paragraph ranking algorithm. This algorithm uses the eigen analysis approach to decide the importance of the paragraph by assigning a rank to each paragraph. Finally the paragraph with the highest rank is identified as the most significant paragraph.

This paper is structured as follows. Section II discusses the current state of research in the area of paragraph ranking. While Section III presents the proposed system architecture, section IV describes the technical details of the proposed paragraph ranking algorithm. The implementation details are explained in the section V. The evaluation process and the results obtained are discussed in Section VI. Finally, section VII concludes this paper.

2. Related Work

Herbert S. Wilf (2002) had mentioned that the idea of using the eigen vector to do ranking is due to Kendall and Wei in the 1950's[4]. This work quantified the importance of web pages using a system of equations which was transformed to a $n \times n$ matrix whose entries were 1 if a link is present between the web pages and 0 otherwise. The importance of each web page was measured by a set of eigen vectors of this matrix. The web page corresponding to the largest entry in that eigen vector was identified as the most important.

Kurt Bryan and Tanya Leise (2006) had analyzed the page rank formula that ranks the importance of web pages according to an eigen vector of a weighted link matrix in their work[6]. They described that the importance scores of each page was computed based on the number of back links for that page. The web page collection was modeled using a directed graph, where an edge from node A to node B indicated a link from page A to page B. In this approach the importance scores for the whole web page collection was defined using a weighted link matrix.

The paragraph ranking algorithm proposed in this work was inspired by the above works.

Mihalcea Rada and Tarau Paul (2004) had discussed the use of TextRank - a graph-based ranking model for text processing. They had mentioned that graph-based ranking algorithm is a way of deciding the importance of a vertex within a graph. To apply this algorithm to natural language texts, they had represented the text as a graph that interconnects text entities with meaningful relations. And the text entities could be of various sizes like words, sentences or others, depending on the application. Also they had compared various graph ranking methods like PageRank and HITS that were proposed to analyze the web page importance[7].

Heng-Hui Liu et al. (2010) and Jung-Hsien Chiang et al. (2011) had proposed a system that recommends significant paragraphs to user. They had devised a measure to evaluate the relevance of a paragraph with respect to a sentence in the abstract (abstract sentence). In this work they had taken the advantage of full text article and the abstract content for ranking paragraphs. The significant paragraph was recommended based on the relevance score assigned to each paragraph[3].

The proposed paragraph ranking algorithm uses a ranking strategy that differs from the existing paragraph ranking approaches. They had evaluated paragraph relevance based on the importance of a paragraph with respect to an abstract sentence. This proposed ranking strategy utilizes the inter-paragraph correlation underlying the text document for measuring their significance. The correlation between the paragraphs was measured by extracting the semantic similarity between them. Further, through the eigen analysis each paragraph was assigned a rank to estimate the importance of paragraphs.

3. Proposed System Architecture

Fig.1. illustrates the system architecture of the proposed paragraph ranking system. The input to the proposed system is a text document and it suggests the most significant paragraph as the output. The most significant paragraph is identified by ranking the paragraphs of the text based on their importance using eigen analysis. The paragraph-paragraph correlation is utilized to compute the paragraph rank values.

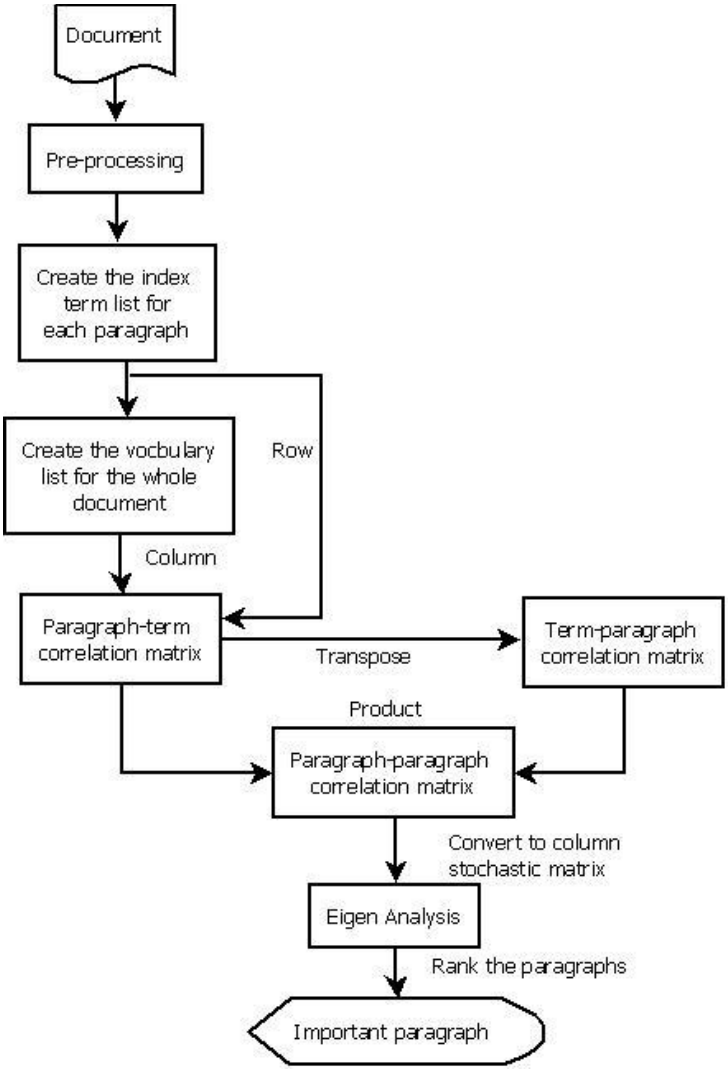

Fig. 1. Proposed system architecture.

4. Methodology

The proposed system takes a text document, D as input and analyzes it to rank the paragraphs according to their importance. The following are the different stages involved in finding the most significant paragraph, P from the whole document D.

4.1. Pre-processing

The full-text document D, was split into paragraphs p_i, $1 < i \leq N$ where N is the total number of paragraphs in the document D. The text was filtered by removing the stop words from it. An index term list K_i was created for each paragraph p_i. Also a vocabulary list $V_d = k_1, k_2, ..., k_M$ was created, where k_j for $1 \leq j \leq M$ are the distinct terms from the whole document. The pre-processing tasks like the tokenization and stop-words removal were done for an effective index term extraction.

A document D can be viewed as a collection of N paragraphs p_i represented in the form of vectors, such that $D = p_1, p_2, ..., p_N$. Each paragraphs can be viewed as an index term list $K_L = k_1, k_2, ..., k_L$, where L is the total number of index terms in paragraph p_i. Thus, the occurrence of the term k_j in paragraph p_i establishes a relation between p_i and k_j. This relation can be quantified using a paragraph-term matrix P, an $N \times M$ matrix whose rows represent the M terms (dimensions) of the N columns, each of which corresponds to a paragraph[2].

Each entry of the matrix P was assigned a non binary weight $0 \leq w_{i,j} \leq 1$, associated with the pair (p_i, k_j). This weight was generated using the tf-ipf weighting technique analogous to the tf-idf weighting technique. The *tf factor* denotes the frequency of a term k_j, in the document D. And the *ipf factor* or inversed paragraph frequency refers to the inverse of the frequency of a term k_j in the collection of paragraphs p_i. The *ipf factor* is analogous to the *idf factor* or the inversed document frequency.

The ipf factor is given by,

$$ipf = \log \frac{N}{n_j}, \tag{1}$$

where N is the total number of paragraphs in the document D and n_j is the number of paragraphs in which the term k_j appears.

The weights according to tf-ipf weighting scheme were calculated using,

$$w_{i,j} = tf \times ipf. \tag{2}$$

Thus the paragraph-term matrix P was created with entries as $w_{i,j}$. The weight $w_{i,j}$ expresses the significance of the term k_j to the paragraph collection.

4.2. Inter-paragraph Correlation

A correlation matrix can be used to represent the semantic similarity or correlation between two paragraphs. Thus a paragraph-paragraph correlation matrix C was generated for measuring the inter-paragraph correlation using:

$$C = P \times P^T. \tag{3}$$

Each entry $c_{i,j}$ of the matrix C expresses the correlation between the paragraphs p_i and p_j. The correlation value $c_{i,j}$, quantifies the semantic similarity between the paragraphs p_i and p_j. Thus, the inter paragraph correlation was captured using the paragraph-paragraph correlation matrix.

In order to enable the application of ranking algorithm to the document D, it has to be represented as a graph. The inter-paragraph correlation was extended to model the document D as a fuzzy-graph based document model. In this model, the document D was represented as a fuzzy graph $G : (\sigma, \mu)$, where σ is a fuzzy subset of a set S and μ is a fuzzy relation on σ. Here the nodes(σ) of the graph G indicated text entities i.e., the paragraphs p_i of the document D. The relations for inter connecting the paragraphs p_i were extracted from the $N \times N$ correlation matrix C. The edges were labeled using the correlation value $c_{i,j}$, where $c_{i,j} \in [0, 1]$. The weight of the edge expressed the fuzzy relation between the paragraphs p_i and p_j in the document D. Thus, using the correlation matrix C, a fuzzy graph based document model was generated. The paragraph ranking algorithm is explained in detail in the following section.

4.3. Paragraph Ranking

To find the most significant paragraph from the fuzzy graph-based document model, the eigen vector centrality of the graph is to be computed. Thus, the paragraph rank values were computed by performing the eigen analysis of the paragraph-paragraph correlation matrix C. For an $N \times N$ square matrix C and a vector $\vec{x}$ that is not all zeros, the values of λ satisfying $C\vec{x} = \lambda\vec{x}$ are called the eigenvalues of C. The N-vector $\vec{x}$ satisfying the above equation for an eigenvalue λ is the corresponding right eigen vector. The left eigen vectors of C are the M-vectors y such that, $y^T C = \lambda y^T$. The eigen vector corresponding to the eigenvalue of largest magnitude is called the principal eigen vector.

The PageRank algorithm, a variant of eigen vector centrality, uses the theory of Markov chain to explain the surfer′s random walk through web pages. A Markov chain is characterized by an $N \times N$ transition probability matrix P each of whose entries are in the interval [0, 1]; the entries in each column of P add up to 1. A matrix with non-negative

entries that satisfies Markov property is known as a *stochastic matrix*. A key property of a stochastic matrix is that it has a principal left eigenvector corresponding to its largest eigenvalue, which is 1[2].

The correlation matrix C obtained was converted into a column stochastic matrix using equation Eq.(4):

$$M = (1 - m)C + mS \tag{4}$$

where $0 \leq m \leq 1$, (here m=0.15)
C denotes an $N \times N$ paragraph-paragraph correlation matrix,
S denotes an $N \times N$ matrix with all entries = $\frac{1}{N}$ and
$M_{N \times N}$ denotes the stochastic matrix equivalent to C[6].

Now the eigenvalues of the transformed correlation matrix M were calculated. The principal eigen vector corresponding to the largest eigenvalue, (i.e., 1) indicates the score of the paragraphs. The paragraph p_i corresponding to the largest entry in the principal eigen vector is identified as the most significant paragraph.

5. Implementation

The proposed paragraph ranking system was implemented in two stages. Algorithm 1 explains the steps involved in the pre-processing stage.

Algorithm 1 Algorithm for pre-processing document D

Require: Document D
Ensure: Index term list K_l of each paragraph p_i and vocabulary list V_d of the document D.

1: Read the document D, and split into paragraphs p_i,where $1 < i \leq N$ and N is the total number of paragraphs in D.
2: Filter the text using stop word removal and identify the index terms.
3: Return the list of index terms K_l, for each paragraph p_i.
4: Identify the vocabulary list V_d for the entire document.

The steps involved in ranking the paragraphs are explained in Algorithm 2. In this stage, the inter-paragraph correlation is computed, which is used for computing paragraph ranks. The algorithm ranks the paragraphs based on the ranks and returns the most significant paragraph.

Algorithm 2 Algorithm for paragraph ranking

Require: N Paragraphs p_i, $1 < i \leq N$ of the document D.
Ensure: Paragraph P_{imp} that is most significant in the document D.

1: Compute tf and ipf for each word v_i in V_d using equation (1).
2: Compute the paragraph-term correlation matrix P with entries $w_{i,j} = tf \times ipf$.
3: Compute the inter-paragraph correlation using the equation (3).
4: Convert C into column stochastic using the equation (4).
5: Compute all the eigval($C_{N \times N}$).
6: Compute the eigvect($C_{N \times N}$) for eigval($C_{N \times N}$) = 1.
7: Return the paragraph P_{imp} corresponding to the largest entry in eigvect($C_{N \times N}$) as the most significant paragraph.

The evaluation of the system is presented in the following section.

6. Evaluation and Results

6.1. Data set

The system was evaluated over the DUC 2001 data set from DUC (Document Understanding Conferences). The data set contains trial, training, and test data from DUC 2001. The test data consist of a randomly selected sample for

each of the 10 NIST assessors of 30 document sets. Each set contained 10 documents on an average, per-document summaries, and multi-document summaries. The guidelines for the task to generate a generic summary of a document automatically were followed to produce the proposed system results. The per-document summaries for each document were of an approximate length of 100 words. And the summaries comprised of two type - original (selector and summarizer is the same person) and duplicate summaries (the selector is a different person from the summarizer)[8]. These human generated are considered as the reference for evaluating the quality of the proposed system generated result.

6.2. Evaluation Metric

The proposed system generated the most significant paragraph that is assumed to convey the relevant information in the text document. The system also ranked all the paragraphs in the text based on their significance. Thus the significant paragraph generated by the system was assumed to be the summary of the text document. Based on this assumption, the quality of the result produced by the proposed system was evaluated using ROUGE evaluation metric.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics and a software package used for evaluating automatic summarization. The metric compares an automatically produced summary against a reference or a set of reference (human-generated) summary to determine the quality of the summary. The measure is computed by counting the number of overlapping words between the automatically generated summary to be evaluated and the reference summaries created by humans. ROUGE-N, ROUGE-L, ROUGEW, ROUGE-S and ROUGE-SU are the available ROUGE measures. ROUGE-N is an N-gram based co-occurrence statistics that is found to be highly correlated with human evaluations[1].

ROUGE-1 (unigram), a metric of informativeness and ROUGE-L (Longest Common Subsequence), metric of fluency and grammaticality are hence considered as the evaluation metrics for the proposed system.

6.3. Results

Table 1. lists a sample of the ROUGE-1 scores of the different documents with 95 % confidence interval. It was observed that for 70% of the text documents the precision values were high when compared to the recall values. The high precision value conveys that the result generated contains more significant information than insignificant information with respect to the reference summaries. The varying paragraph length of the different documents affected the ROUGE scores. Hence the documents with short paragraphs produced a low recall and high precision.

Table 1. ROUGE-1 Scores (95%-conf.int.) for the most significant paragraph

Document ID	Recall	Precision	F-measure
1	R:0.60688	P:0.53233	F:0.56717
2	R:0.48238	P:0.40556	F:0.44065
3	R:0.42432	P:0.36009	F:0.38958
4	R:0.42955	P:0.37500	F:0.40043
5	R:0.41628	P:0.42619	F:0.42118
6	R:0.41457	P:0.51562	F:0.45961
7	R:0.38608	P:0.59416	F:0.46803
8	R:0.39052	P:0.48596	F:0.43304
9	R:0.36629	P:0.50938	F:0.42614
10	R:0.40330	P:0.41106	F:0.40714
11	R:0.16596	P:0.60938	F:0.26087
12	R:0.15268	P:0.61486	F:0.24462
Average	R:0.35962	P:0.46726	F:0.39496

Table 2. lists a sample of the ROUGE-L scores of the different documents. ROUGE-L scores being the metric for fluency, was observed to be relatively good for the significant paragraph. The result produced by the proposed system was a paragraph from the text document, which obviously would offer more readability than extractive summaries. The ROUGE-L scores also reflected a high precision for ($\approx$ 75 %) of the text documents.

Table 2. ROUGE-L Scores (95%-conf.int.) for the most significant paragraph

Document ID	Recall	Precision	F-measure
1	R:0.57985	P:0.50862	F:0.54190
2	R:0.47577	P:0.40000	F:0.43461
3	R:0.41081	P:0.34862	F:0.37717
4	R:0.40704	P:0.50625	F:0.45126
5	R:0.36498	P:0.56169	F:0.44246
6	R:0.38208	P:0.38942	F:0.38572
7	R:0.38608	P:0.59416	F:0.46803
8	R:0.39052	P:0.48596	F:0.43304
9	R:0.36629	P:0.50938	F:0.42614
10	R:0.40330	P:0.41106	F:0.40714
11	R:0.16596	P:0.60938	F:0.26087
12	R:0.15268	P:0.61486	F:0.24462
Average	R:0.32512	P:0.42609	F:0.34870

Table 3. and Table 4. list a sample of ROUGE-1 and ROUGE-L scores of the different documents. It was observed that for ($\approx$ 30%) of the text documents, the paragraphs ranked second or third by the proposed system was having relatively higher ROUGE scores than that of the most significant paragraph. The recall and precision values followed a similar pattern as in the case of the significant paragraph. However, it could be observed that in order to have a fair understanding about the information in a text document, it is not only required to read the most significant paragraph but also the second and third ranked paragraph.

Table 3. ROUGE-1 and ROUGE-L scores for the second ranked paragraph.

ROUGE-1 Scores (95%-conf.int.)				ROUGE-L Scores (95%-conf.int.)		
Doc.ID	Recall	Precision	F-Measure	Recall	Precision	F-Measure
1	R:0.38177	P:0.69196	F:0.49206	R:0.43188	P:0.38182	F:0.40531
2	R:0.40639	P:0.52976	F:0.45995	R:0.40969	P:0.39407	F:0.40173
3	R:0.41850	P:0.40254	F:0.41036	R:0.39269	P:0.51190	F:0.44444
4	R:0.32731	P:0.50347	F:0.39671	R:0.37192	P:0.67411	F:0.47936
Average	R:0.29462	P:0.40900	F:0.31681	R:0.28527	P:0.39354	F:0.30557

Table 4. ROUGE-1 and ROUGE-L scores for the third ranked paragraph.

ROUGE-1 Scores (95%-conf.int.)				ROUGE-L Scores (95%-conf.int.)		
Doc. ID	Recall	Precision	F-Measure	Recall	Precision	F-Measure
1	R:0.46272	P:0.40179	F:0.43011	R:0.35910	P:0.52941	F:0.42793
2	R:0.34989	P:0.49051	F:0.40844	R:0.35135	P:0.55000	F:0.42878
3	R:0.36908	P:0.54412	F:0.43982	R:0.43959	P:0.38170	F:0.40860
4	R:0.35872	P:0.56154	F:0.43778	R:0.19493	P:0.69444	F:0.30441
Average	R:0.28939	P:0.43882	F:0.31508	R:0.27062	P:0.41447	F:0.30449

The results obtained indicate that the most significant paragraph suggested by the paragraph ranking module has a fair readability and conveys majority of the information in a text document. Apart from the significant paragraph, when the second or third ranked paragraphs are combined together results in a relatively good text summary.

7. Conclusion

This work proposes a novel paragraph ranking method that ranks the paragraphs in a text document according to their importance. The proposed paragraph ranking algorithm utilizes the inter paragraph correlation to determine

the paragraph ranks. The proposed algorithm exploits the content of the document and hence can identify the most significant paragraph even from non technical documents. The most significant paragraph of the text document was identified on the basis of eigen analysis under the assumption that it would convey the significant information in the text document. The results produced by the system were evaluated using the DUC 2001 data set and the ROUGE evaluation metric. The evaluation results show that the significant paragraph suggested based on eigen analysis had a good readability and covered a good amount of information contained in the text document.

References

1. Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Spain; 2004.
2. Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press: Cambridge; 2009.
3. Heng-Hui Liu, Yi-Ting Huang, Jung-Hsien Chiang. A study on paragraph ranking and recommendation by topic information retrieval from biomedical literature. *In proceeding of the International Conference on Computer Symposium (ICS)*. p.859–864; 16-18 Dec. 2010.
4. Herbert S. Wilf. Searching the web with eigenvectors. *UMAP Journal*. Vol.23; Issue 2; p.101; June 2002.
5. Jung-Hsien Chiang, Heng-Hui Liu, Yi-Ting Huang. Condensing biomedical journal texts through paragraph ranking. *Bioinformatics*. Vol.27; No.8; p.1143–1149; 2011.
6. Kurt Bryan, Tanya Leise. The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review*. Vol.48; Issue 3; p.569–581; 2006.
7. Mihalcea Rada, Tarau Paul. TextRank: Bringing Order into Texts. *Proceedings of EMNLP–04 Empirical Methods in Natural Language Processing*. July 2004.
8. *http*: *//www.nlpir.nist.gov/projects/duc/past_duc/duc*2001/ *data.html*. Last visited on: August 2014.