Comparison of Document Index Graph Using TextRank and HITS Weighting Method in Automatic Text Summarization

Fadhlil Hadyan¹, Shaufiah² and Moch. Arif Bijaksana³
School of Computing, Telkom University, Bandung, Indonesia

¹fadhilhadyan0@gmail.com, ²shaufiah@telkomuniversity.ac.id, ³arifbijaksana@telkomuniversity.ac.id

Abstract. Automatic summarization is a system that can help someone to take the core information of a long text instantly. The system can help by summarizing text automatically. There's already many summarization systems that have been developed at this time but there are still many problems in those systems. In this final task proposed summarization method using document index graph. This method utilizes the PageRank and HITS formula used to assess the web page, adapted to make an assessment of words in the sentences in a text document. The expected outcome of this final task is a system that can do summarization of a single document, by utilizing document index graph with TextRank and HITS to improve the quality of the summary results automatically.

1. Introduction

The digital era enables a lot of people to obtain article or news very easily through internet using their computer like pc or tablet [1]. Especially, we use these devices constantly in our entire day and a lot of article is available in the internet. This results in a large number of text data generated every day. Although we have a large number of text, utilization of this data is very low, especially in a long text. It’s not easy to obtain information from a long article with a short time, thus we need to take more time to read the text. It is necessary to automatically extract core information from an article, since it will reduce the time we need to obtain the information.

Automated text summarization aims to generate summary from a document or new in this research, reducing length of a news while keeping the same core information as the original news. The goal of document summarization is to expressing the original document with fewer words [2]. There are two types of document summarization, which is extractive and abstractive [3]. Extractive based summarization identifies and pick important sentence from document and combine the sentences into summary. Abstractive summarization attempt to generating new sentence, but currently this type off summarization is still difficult to implement. Majority of automated summarization system is still using extractive method [4]. This research is going to develop an extractive automated summarization system using document index graph.
Document index graph is a method to represent document as graph. With this representation a sentence will not be scored based on words. Sentence scoring is based on similarity of phrase [5]. Since phrase will be represented by connected nodes in graph. This ability to represent phrase is the advantage of document index graph. Graph scoring algorithm, TextRank and HITS, will be used to score the graph produced by DIG method. this scored nodes will be used to score sentence and identify important sentence.

2. Theoretical Review
Sentence scoring is an attempt to identify important sentences of a document. Sentence scoring can be divided into two process, graph construction and graph scoring. Graph scoring can utilize several algorithms like HITS or lexrank. This paper is using TextRank and HITS to score graph constructed by document index graph. Sentence score will determine by the score of each word contain in that sentence. Sentence with high score will be included as a result summary. Graph construction is using document index graph algorithm to create graph from word in a document.

2.1. Document index graph
Document index graph constructing graph from words inside a document, stopword will not be include in graph construction. DIG is a directed and weighted graph consist of nodes and edges. Nodes represent word in document, where edges represent occurrence of two words in the document [6]. As a weighted graph, each edges have its own weight. Edges weight is the frequency of occurrence of the two words in sequence.

In single document summarization, graph is constructed for each document. Since we assume there is no correlation between document we construct graph for single document. In each document index graph iterate through each word. The construction of document index graph is following this two rules:

- If a word never occurred before, then create new node (node A) and draw edge from node B (node in previous iteration) to node A, if node A is not the first word in the sentence. the weight of new edge is 1.
- If a word has occurred before, find the node that containing current word (node A) and draw edge from node B (node in previous iteration) to node A, if node A is not the first word in the sentence. If node A and B already connected by an edge, increment its edge by 1.

![Figure 1. Example of DIG result](image)
Document index graph method is able to score sentence based on words or phrase inside the sentence [7]. In figure 1, phrase is represented by two connected nodes with high edge weight. A phrase is two nodes that occurred in sequence multiple times. with this representation, phrase can be stored inside graph alongside all other word, thus there’s no need to store phrase information separately and reduce resource requirement [5].

2.2. TextRank
TextRank is a graph scoring algorithm, TextRank attempt to determine score for all nodes based on information contained in the graph itself [8]. TextRank algorithm in based on web-site scoring algorithm PageRank, the original page rank is used to handle unweighted graph. since TextRank is going to score weighted graph, the original formula is changed to enable TextRank handles weighted graph.

Formula (1) [8] is an adapted PageRank formula to integrate edges weight.

\[
PR(A) = (1 - d) + d \sum_{i=1}^{n} \frac{PR(T_i)}{C(T_i)}
\]

- \(PR(x)\) = weight of nodes
- \(n\) = number of nodes in graph
- \(C(x)\) = number of out edge from node \(x\)
- \(D\) = damping factor, a value between 0 and 1, dampling factor is usually set to 0.85 [9].

In formula (1), score of PR is determined by other PR score, thus PR score need to be initialize with a score, the initial score can be any number except for 1, since PR score will converge to 1 [10]. Number of iteration needed to complete TextRank algorithm is define by a convergence threshold, the threshold used in this paper is 0.0001 [8]. TextRank algorithm will stop after error is lower than the threshold, error in the difference between score of a same node in different iteration [7].

formula (2) [7] describe how error will be calculated:

\[
e = PR^{i+1} - PR^i
\]

2.3. HITS
HITS or hyperlinked induced topic search is another graph scoring algorithm, this algorithm using two score, authority and hubs, instead of a single score like in TextRank [11]. HITS algorithm will determine nodes with high authority (node with large number of income edges) and node with high hub (node with large number of outgoing edges). There are two different formulas that will used to calculate each score.

Formula (3) and (4) [11] is HITS formula used to score graph:

\[
HITS_A(V_i) = \sum_{V_j \in V_i} HITS_H(V_j)
\]
\[ HITS_H(V_i) = \sum_{V_j \in \text{in}(V_i)} HITS_A(V_j) \]  

- HITS_A(x) = nodes authority score
- HITS_H(x) = nodes hub score

HITS is an iterative algorithm formula (3) and (4) will use in sequence, each node will have starting assigned value. Similar to TextRank, HITS calculate each score until convergence below a certain threshold [12].

2.4. Rouge-n
The output result of a system need to be evaluate to analyse the system quality and improvement that could be made. ROUGE or recall-oriented understudy for gisting evaluation is an evaluation method that measure text similarity based on n-gram similarity between text [13]. ROUGE is a unit to compared similarity between system generated summary and reference summary or the gold standard. Formula (3) [13] describe how ROUGE calculate recall value of system generated summary by number of n-gram found in both summary (system and reference) divided by total number of n-gram in reference summary.

\[
\text{ROUGE} - N = \frac{\sum_{CERSS} \sum_{gram_n \in C} Count_{match}(gram_n)}{\sum_{CERSS} \sum_{gram_n} Count(gram_n)}
\]  

3. System Architecture

News dataset will process using python programming language, and evaluate using ROUGE toolkit. News file will first be preprocessed to clean text before transformed into graph. the news file will split into list of sentence and each sentence will have tokenized in to list of words. Stopword removal applied in the list to remove common words that not significance in news (e.g. a, is, it, etc.). stemming
process will remove suffix (e.g. -ing, -ly, -er) to transform word to its root. The result of stemming will allow different word with same root score the same.

Preprocessed text will be transformed into graph. Each sentence will iterate, thus last word in the sentence will not connected to first word in the next sentence. This process allows possibility of dangling nodes, which is a node that doesn’t have out edge. TextRank algorithm cannot handle dangling nodes by itself, thus an extra process is needed to handle this condition. Dangling nodes can be handled by connecting it to all other nodes in graph [14]. The result graph will be transformed again into adjacency matrix to further scored using TextRank algorithm. Scored word then use to score each sentence. Summary will generate by picking sentence with highest score.

The system generated summary will evaluated using rouge-2. System summary will compare to reference summary from news. Average recall of each summary will become the evaluation score of the system.

4. Testing and Analysis

4.1. Dataset
The dataset used in this research from MultiLing 2015 also used in [15], this research use 30 document with Indonesian language from the dataset. The original and reference text is stored in txt format in two separate folders. Original and reference document using a same file name with slight difference after the filename, ‘_body.txt’ for original text and ‘_summary.txt’ for reference summary. Each reference summary has average range at 218.33 words.

4.2. Evaluation Result and Analysis
The system will be evaluate using rouge-1, also used in [15], which resulting recall, precision and f-measure. Evaluation will perform on 3 scenarios with 3 different compression rate, 30%, 40%, and 50%. The result of ROUGE-1 evaluation on three scenarios is show on table 1.

| Summary Length | Recall      | Precision | F-measure |
|----------------|-------------|-----------|-----------|
| 50             | 0.074069    | 0.332153  | 0.120741  |
| 100            | 0.147472    | 0.327876  | 0.202573  |
| 150            | 0.209757    | 0.313522  | 0.250186  |
| 200            | 0.265075    | 0.297331  | 0.279062  |
| 250            | 0.309068    | 0.277654  | 0.291242  |
Table 3. HITS Evaluation result.

| Summary Length | Recall   | Precision | F-measure |
|----------------|----------|-----------|-----------|
| 50             | 0.078226 | 0.349438  | 0.127522  |
| 100            | 0.14939  | 0.331628  | 0.205178  |
| 150            | 0.217959 | 0.320518  | 0.258361  |
| 200            | 0.273328 | 0.302407  | 0.28592   |
| 250            | 0.322984 | 0.287407  | 0.302845  |

Figure 3. Scoring Algorithm Comparison

Based from table-2 and table-3, f-measure of the system is increasing as the number of words, the result is consistent between two systems. While recall following f-measure trend, precision have the opposite trend. Precision of the systems decrease as the length of the summary increased, this follow the usual precision-recall trend. The highest f-measure from both comes from summary with 250 word. The overall result is pretty low compared to the system mentioned in [15]. The system developed in this research is general text summarization system, while the reference summary is topic-focused summaries. The best f-measure score coms from HITS system with f-measure of 0.302845, slightly higher that TextRank best result of 0.291242.

Figure 3 show the comparison between TextRank and HITS on all summary length. The trend consist HITS have the highest score in all summary length, but not a large difference. The smallest difference is in 100 words summary with difference under 0.003.
5. Conclusion

From this research result, it can be concluded:

1. Summarization using document index graph with TextRank and HITS resulting f-measure score of 0.291242 and 0.302845 respectively. Furthermore, HITS algorithm resulting a slightly higher score than TextRank.

2. The average f-measure score of summary highly dependent on the length of the summary.

6. References

[1] K. LU and J. HOLCOMB, “Digital News Audience: Fact Sheet,” Pew Research Center, 15 June 2016. [Online]. Available: http://www.journalism.org/2016/06/15/digital-news-audience-fact-sheet/. [Accessed 11 November 2016].

[2] L. Marina and L. Matk, “Graph-Based Keyword Extraction for Single-Document Summarization,” Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization, p. 8, 2008.

[3] E. Hovy and C.-y. Lin, “AUTOMATED TEXT SUMMARIZATION,” p. 18, 1998.

[4] G. R. Radev, E. Hovy and K. Mckeown, “Introduction to the special issue on summarization,” Journal Computational Linguistics - Summarization, vol. 28, no. 4, pp. 399-408, 2002.

[5] K. M. Hammouda and M. S. Kamel, “Efficient Phrase-Based Document Indexing for Web Document Clustering,” p. 18, 2004.

[6] Q. Zhou, L. Sun and J.-Y. Nie, “IS_SUM: A Multi-Document Summarizer based on Document Index,” 2005.

[7] M. a. Jinnah, “Analisis dan Implementasi Peringkasan Teks Otomatis Dokumen Tunggal Berbahasa Indonesia Menggunakan Document Index Graph,” 2011.

[8] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Texts,” Association for Computational Linguistics, 2004.

[9] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Seventh International World-Wide Web Conference, 1998.

[10] I. Rogers, “The Google Pagerank Algorithm,” IPR Computing Ltd, [Online]. Available: http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm. [Accessed 11 June 2016].

[11] J. M. Kleinberg, “A Authoritative Sources in a H y p erlinked Env ironment,” Journal of the ACM (JACM), vol. 46, pp. 604-632, 1999.

[12] R. Mihalcea and P. Tarau, “A language independent algorithm for single and multiple document summarization,” 2005.

[13] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” Text summarization branches out: Proceedings of the ACL-04 workshop, vol. 8, 2004.

[14] E. Yan and Y. Ding, “The effects of dangling nodes on citation networks,” 2011.

[15] M. Vicente, O. Alcon and E. Lloret, “The University of Alicante at MultiLing 2015: approach, results and further insights,” 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2015.