Jurnal SimanteC
Vol. 9, No. 1 Desember 2020
P-ISSN : 2088-2130
E-ISSN : 2502-4884

RINGKASAN DOKUMEN ILMIAH BERBASIS REPRESENTASI DATA *GRAPH* MENGGUNAKAN UKURAN *CENTRALITY*

Mulaab[1)]
[1]Prodi Teknik Informatika, Fakultas Teknik,Universitas Trunojoyo
Jl. Raya Telang, PO BOX 2 Kamal, Bangkalan
E-mail : [1]mulaab@trunojoyo.ac.id

ABSTRAK

Saat ini, informasi banyak tersedia di internet, sangat diperlukan suatu teknik untuk mendapatkan informasi dengan cepat dan paling efisien. Diantaranya terdapat sumber informasi ilmiah yang berasal dari teks yang tersedia sangat banyak di internet. Oleh karena itu diperlukan teknik dan mekanisme yang baik untuk mengekstrak informasi yang paling relevan darinya. Teknik peringkasan dokumen ilmiah adalah salah satu metode untuk mengkompresi dari isi dokumen yang besar menjadi teks yang lebih pendek. Ringkasan teks yang dihasilkan mengandung pokok-pokok makna dan gagasan yang ada dalam dokumen asli. Ringkasan teks atau dokumen ilmiah berdasarkan ekstraksi adalah memilih sebagian kalimat yang memiliki relevansi tinggi (peringkat) dari dokumen berdasarkan fitur kata dan kalimat tersebut selanjutnya menggabungkan kalimat-kalimat yang dipilih untuk menghasilkan ringkasan dokumen. Pada makalah ini menggunakan model pemeringkatan kalimat penting berdasarkan ukuran *centrality* yaitu pentingnya simpul-simpul pada data *graph*. Kalimat direpresentasikan sebagai data berbasis *graph* dengan simpul-simpul dari suatu *graph*.

Kata kunci : *Ringkasan Text Otomatis, Centrality, Graph Analysis, TF-IDF*

ABSTRACT

Today, information is widely available on the internet, a technique is needed to get information quickly and most efficiently. Among them, there are sources of information derived from texts that are widely available on the internet. Therefore we need good techniques and mechanisms to extract the most relevant information from it. The text summarization technique is a method for compressing large document contents into shorter text. The resulting text summary contains key meanings and ideas contained in the original document. Text summary based on extraction is to select part of the sentence which has high relevance (ranking) from the document based on the word and sentence feature, then combine the selected sentences to produce a document summary. This paper uses a ranking model for important sentences based on the measure of centrality, namely the importance of the nodes in the data graph. Sentences are represented as graph-based data with the vertices of a graph

Keywords: *Text Summarization, Centrality, Graph Analysis, TF-IDF*

PENDAHULUAN

Peringkasan teks otomatis adalah tugas menghasilkan ringkasan dokumen dengan mempertahankan informasi utama dan arti secara keseluruhan [1]. Terdapat dua jenis kelompok teknik dalam melakukan peringkasan teks. Yaitu peringkatan teks ektraktif (*Extractive text astraktif summarization*) dan peringkasan teks *abstraktif* (*Abstractive text summarization*)

Peringkasan teks *ektraktif* (*Extractive text summarization*) Peringkasasn teks *ektraktif* merupakan teknik peringkasan teks dengan mengidentifikasi dan mengektrak kata-kata penting, *phrase* atau kalimat-kalimat dari dokumen dan menjadikannya menjadi ringkasan atau kesimpulan dari dokumen. Sehingga ringkasan yang dihasilkan adalah bagian dari kalimat dari dokumen aslinya [2]. Sebagian besar penelitian peringkasan dokumen menggunakan metode peringkasan *ektraktif*.

Peringkasan teks *abstraktif* memungkinkan kita untuk membuat ringkasan seperti cara orang membuat ringkasan. Secara umum model peringkasan *abstraktif* lebih jelek dari model peringkasan *ektraktif*. Hasil ringkasannya kadangkala sangat menyimpang dengan isi dari dokumen aslinya. Model peringkasan ini juga memungkinkan menggunakan kata-kata yang berbeda dengan kata-kata dalam dokumen aslinya. Dengan gaya lebih mirip dengan gaya orang membuat ringkasan, maka model ini sangat menjanjikan untuk topik kecerdasan buatan.

METODE

Model Peringkasan Teks

Terdapat tiga model yang digunakan untuk ringkasan teks yaitu *statistical frequency computation models* (TFIDF etc.), model berbasis *graph*, dan pendekatan berbasis pembelajaran mesin (*machine learning models*)

1. Model Statistik

Metode ini didasarkan pada asumsi bahwa pentingnya dari kata atau kalimat bergantung pada banyaknya kemunculan kata atau kalimat tersebut dalam dokumen. Ini berarti bahwa pendekatan klasik ini mengabaikan kontek dan fitur *leksikal* dari teks. Oleh karena itu, pendekatan ini hanya dapat digunakan pada jenis ringkasan *ektraktif (extractive summarization)*

2. Model berbasis Graph

Pada model berbasis *graph*, *graph* merupakan representasi dari suatu dokumen dengan simpul-simpul *graph* adalah kalimat-kalimat dalam dokumen. Bobot sisi dari *graph* menyatakan kemiripan (*similarity*) antara kalimat dengan kalimat dalam suatu dokumen. Model berbasis graph masih ada keterbatasan dalam hal pemahaman leksikal. Sehingga model ini hanya dapat digunakan untuk melakukan peringkasan *ektraktif*.

3. Model dengan Pembelajaran Mesin (Machine Learning)

Saat ini, banyak penelitian yang fokus pada pelatihan mesin pada *Natural Language Processing* (NLP) termasuk peringkasan dokumen. Model dengan basis pembelajaran mesin untuk melakukan peringkasan dengan model pembelajaran mesin dengan mengintegrasikan teknik *unsupervised deep neural network* dan pendekatan *word embedding*[3]. Model pengembangan peringkasan teks untuk mengulas opini [4-6]

Graphs Pada Ringkasan Teks

Suatu *graph* $G(V,E)$ adalah struktur matematika yang digunakan untuk merepresentasikan relasi berpasangan antara objek dengan objek. *Graph* memiliki dua komponen V *(Vertices)* atau simpul dan E *(Edge)* atau sisi. Simpul (vertices) menyatakan komponen utama dari sistem yang direpresentasikan dan *edge* atau sisi merepresentasikan relasi antara dua simpul *(verteks)*. Untuk

merepresentasikan suatu dokumen menggunakan model berbasis *graph*, butuh tiga hal.

1. Unit dasar dari aplikasi, dalam ringkasan teks adalah kata, kalimat atau mungkin paragraph
2. Tipe relasi antara simpul-simpul untuk menghitung bobot dari sisi dalam ringkasan teks, misal cosine *similarity*.
3. Algoritma pemeringkatan yang digunakan untuk memberi peringkat simpul dalam *graph*

Algoritma PageRank

Internat jaringan yang sangat komplek. Relasi antara laman dengan laman dapat dinyatakan dengan *graph*. Pentingnya suatu simpul adalah dapat dilihat dari keluar masuknya *link* pada suatu simpul. Banyaknya *link* masuk dari suatu laman menyatakan indikasi laman tersebut penting atau berkualitas. Alogritma PageRank menggunakan ide ini untuk memberi peringkat pada laman pada hasil proses pencarian[7]. Algoritma PageRank menjadikan nilai pentingnya suatu laman tertentu bergantung pada nilai pentingnya laman–laman yang menunjuk kepada laman tersebut.

Evaluasi Model

Mengevaluasi model peringkasan dokumen adalah tugas yang sangat sulit karena tidak ada satu ringkasan tunggal yang ideal dari dokumen yang diringkas. Bahkan dalam banyak kasus ditemui hasil ringkasan otomatis tidak disetujui oleh orang *(evaluator)*. Oleh karena itu kita harus membuat asumsi tentang ruang untuk membuat ringkasan yang baik.

1. Mengasumsikan bahwa kesimpulan baik akan mendekati kesimpulan yang dibuat oleh manusia
2. Mengasumsikan bahwa bagusnya ringkasan dapat diukur dengan banyaknya informasi penting yang ada didalam dokumen. (Asumsi ini dasarkan dari definisi peringkasan teks).

ROUGE

Ukuran evaluasi ini adalah yang paling banyak digunakan untuk memberikan skor pada saat mengevaluasi peringkasan teks. Metode *ROUGE (Recall-Oriented Understudy for Gisting Evaluation)* diperkenalkan oleh Chin-Yew Lin in 2004 [8]

$$ROUGE-N = \frac{\sum_{gram_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)} \quad (1)$$

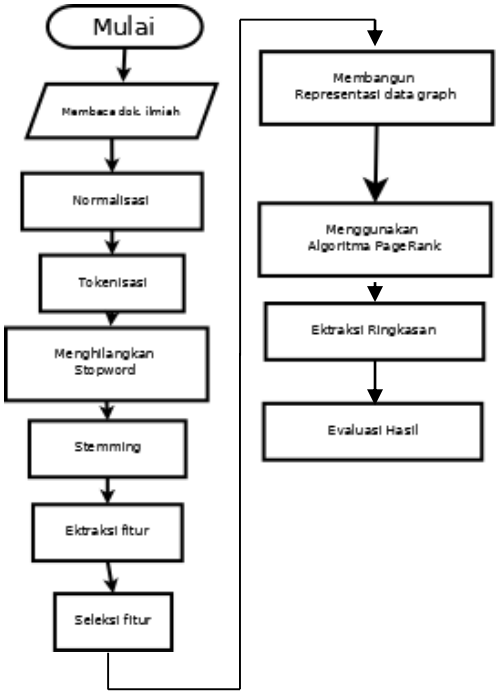
Dimana $gram_n$ adalah pilihan $n-gram$ dan S adalah referensi ringkasan.

Metode Usulan

Pada bagian ini akan menjelaskan pendekatan yang diusukan untuk melakukan ringkasan dokumen ilmiah. Gambar 1 menunjukkan diagram alir dari pendekatan yang diusulkan.

Beberapa proses diatas dapat dijelaskan sebagai berikut:

1. Membaca dokumen ilmiah, adalah tahapan awal sistem dengan menentukan dokumen ilmiah yang akan diektrak ringkasannya
2. Normalisasi adalah proses untuk membuang tanda hubung, tanda baca lainya, angka dan lain-lain
3. *Tokenisasi.* Pada langkah ini dokumen dibagi menjadi beberapa paragraf, kemudian paragraf menjadi kalimat, dan terakhir kalimat menjadi kata-kata



Gambar 1. Pendekatan yang diusulkan

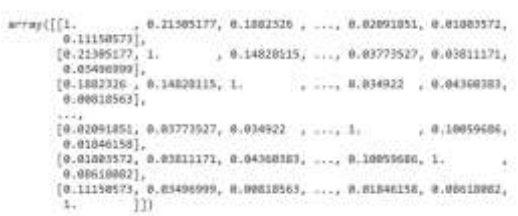
4. Menghapus *stopword*. Menghapus *stopword* adalah mengurangi teks menjadi teks teks yang lebih berguna. Jika tidak menghapus *stopword* maka akan mempengaruhi efisiensi dari proses pembobotan
5. *Stemming*. *Stemmer* digunakan untuk mengekstrak kata dasar dari setiap kata dalam kalimat. Proses ini dapat mengurangi ragam kata dalam dokumen sehingga memperbaiki perhitungan frekuensi kosakata
6. Ektraksi fitur adalah proses membangun fitur dari setiap kalimat dalam suatu dokumen
7. Seleksi fitur adalah proses menentukan fitur, dalam hal ini adalah memilih kata dasar yang relevan (kata dasar penting) yang memberikan karakteristik penting dari suatu dokumen.
8. Membangun representasi *graph*. Pada tahapan ini keterkaitan kalimat dengan kalimat dinyatakan dalam suatu *graph*. Proses diawali dengan membentuk matrik *adjancy* dari kalimat dengan kalimat lainya dalam dokumen ilmiah.
9. Menghitung *Pagerank* simpul-simpul graph yang menyatakan ukuran *centrality* dari pentingnya simpul-simpul *graph*. Dalam hal ini ukuran *centrality* yang digunakan menggunakan algoritman *PageRank Graph* yang menyatakan representasi hubungan kalimat dengan kalimat lainnya pada dokumen ilmiah, akan dihitung nilai *PageRank* setiap simpulnya.
10. Ektraksi ringkasan. Pada tahapan ini adalah memilih kalimat-kalimat tertentu berdasarkan nilai *PageRank* suatu kalimat dengan batas ambang yang telah ditetapkan.
11. Evaluasi hasil adalah untuk menunjutkkan ukuran relevansi kalimat terhadap ringkasan yang telah dibuat.

HASIL DAN PEMBAHASAN

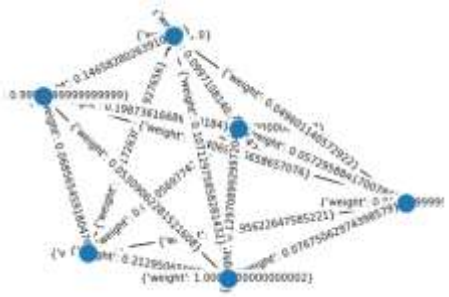
Data dan Diskripsi dokumen Ilmiah

Data yang digunakan dalam penelitian ini adalah data yang berasal dari repositori https://github.com/WING-NUS/scisumm-corpus. Tepatnya pada https://github.com/WING-NUS/scisumm-corpus/tree/master/data/

Data ini terdiri makalah ACL *Anthology*. Setelah melakukan *praproses* terhadap data dokumen maka didapatkan matrik *adjancy*, hubungan kalimat dengan kalimat seperti berikut seperti gambar 2, dan bagian dari proses pembentukan *graph* dari kalimat dengan kalimat yang memiliki bobot dapat dinyatakan dengan gambar 3.



Gambar 2. Matrik *Adjancy* dari kalimat dalam kalimat dokumen



Gambar 3. Bagian Proses Pembentukan Kalimat dan bobot matrik *adjancy*

Kesimpulan Sistem: *Specifically, this study intends to provide the following contributions to the frameworks of the capacity allocation problem based on the optimization methods: Using the theory of rail network and the train and passenger flow space–time network representation respectively, a detailed description for the structure and characteristics of train capacity allocation problem under stochastic demands in the high-speed rail network is presented. The train capacity allocation problem under stochastic demands in the high-speed rail network is formulated as a two-stage stochastic integer programming model in Section 3.In this paper, we propose a stochastic integer programming model for passenger capacity problem under random passenger demands and some strategies for optimizing train timetabling problem in the high-speed rail network by using network optimization techniques"*

Kesimpulan Manual : *As tactical plans of complex rail operations, train timetables are programmed and updated every year or every season because of remarkable variation of passenger demands.Using the theory of rail network and the train and passenger flow space–time network representation respectively, a detailed description for the structure and characteristics of train capacity allocation problem under stochastic demands in the high-speed rail network is presented.The train capacity allocation problem under stochastic demands in the high-speed rail network is formulated as a two-stage stochastic integer programming model in Section.Through implementing on the Beijing-Shanghai high-speed rail network in China, we verify the performance and effectiveness of the proposed methods*.".

Dari dokumen ilimiah yang diproses diperoleh ukuran *ROUGE-L* yang diberikan:
Precision :0.7698309492847855
***Recall** is* :0.6773455377574371
***F Score** is* :0.72063393864

KESIMPULAN

Dari hasil peringkasan teks yang dilakukan dengan menggunakan *PageRank* menunjukkan nilai *F-score* yang tinggi mendekati nilai 1, yang memberikan informasi bahwa Nilai tertinggi untuk skor-F adalah 1, yang berarti bahwa cukup baik dalam peringkasan teks dengan ditunjukkan peringmbangan dengan nilai ***Precision*** dan ***Recall***

DAFTAR PUSTAKA

[1] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, *et al.*, "Text summarization techniques: a brief survey," *arXiv preprint arXiv:1707.02268*, 2017.

[2] J. K. Yogan, O. S. Goh, B. Halizah, H. C. Ngo, and C. Puspalata, "A review on automatic text summarization approaches," *Journal of Computer Science*, vol. 12, pp. 178-190, 2016.

[3] N. Alami, M. Meknassi, and N. En-nahnahi, "Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning," *Expert systems with applications*, vol. 123, pp. 195-211, 2019.

[4] M. Hong and H. Wang, "Research on customer opinion summarization using topic mining and deep neural network," *Mathematics and Computers in Simulation*, 2020.

[5] P. Wu, X. Li, S. Shen, and D. He, "Social media opinion

summarization using emotion cognition and convolutional neural networks," *International Journal of Information Management*, vol. 51, p. 101978, 2020.

[6] A. Abdi, S. Hasan, S. M. Shamsuddin, N. Idris, and J. Piran, "A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion," *Knowledge-Based Systems*, p. 106658, 2020.

[7] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Stanford InfoLab1999.

[8] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74-81.