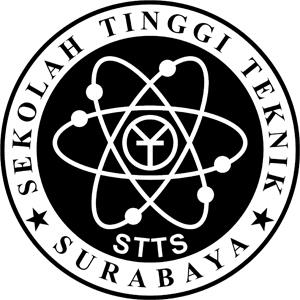
**PROPOSAL**

**Project Mata Kuliah Software Developer Project**



**221116935 - Aldi Afendiyanto**

**221116958 - Geovann Chandra**

**22116936 - Alvin Bernard Wiyono**

**221116947 - Darrell Fiko Alexander**

**INSTITUT SAINS TERAPAN DAN TEKNOLOGI SURABAYA**

**AGUSTUS 2023**

**Proposal Untuk Project Mata Kuliah Software Developer Project Dengan Studi Kasus Website Perusahaan Distributor Barang Primer**

# Latar Belakang

Perangkuman teks secara otomatis (*automatic text summarization*) merupakan proses pengambilan dokumen tekstual, mengekstraksi isinya, dan menyajikan konten yang paling penting untuk pengguna dalam bentuk yang lebih padat dan sesuai dengan kebutuhan pengguna dengan memanfaatkan aplikasi yang dijalankan dan dioperasikan sebagai perangkat lunak (Kogilavani dan Balasubramami, 2010). Ringkasan merupakan teks yang dihasilkan dari sebuah teks atau banyak teks yang mengandung informasi dari teks asli dan panjangnya juga tidak lebih dari setengah baris teks aslinya (Hovy, 2001). Penelitian mengenai perangkuman secara otomatis (automatic text summarization) dengan memanfaatkan berbagai metode dan pendekatan, diawali oleh Luhn sejak tahun 1958.

Perangkuman teks dapat diklasifikasikan menjadi dua kategori, yaitu: ekstraksi (*ekstraktive*) dan abstraksi (*abstractive*). Perangkuman ekstraksi berarti melakukan seleksi kalimat atau frasa dari teks asli yang memiliki nilai tinggi kemudian dikemas menjadi satu agar mendapatkan teks baru yang lebih pendek tanpa merubah makna dari teks asli. Perangkuman abstraksi berarti memanfaatkan metode linguistik dalam memeriksa dan menafsirkan teks asli sehingga kata-kata dan frase yang didapatkan tidak ada di dalam dokumen yang asli. (Suanmali L, Salim N, Salem B, 2009).

Pada saat ini, pengembangan sistem perangkuman tidak hanya satu dokumen saja, melainkan berasal dari banyak dokumen. Perangkuman multi- dokumen menggunakan kategori perangkuman teks ekstraktive ( He dkk, 2008; Wan dan Yang, 2008; Sarkar, 2009; Ge dkk, 2011; Ouyang dkk, 2013). Kalimat- kalimat yang diekstraksi dari dokumen sumber dipilih berdasarkan kriteria tertentu. Pada metode ini juga dituntut mampu menyampaikan konten-konten penting dari dokumen-dokumen sumber.

Ada dua permasalahan yang sering dihadapi saat meringkas suatu dokumen, yaitu: coverage dan saliency. Coverage berarti ringkasan yang memiliki cakupan sebanyak mungkin sedangkan saliency berarti konsep-konsep penting yang ada pada dokumen sumber (Ouyang dkk, 2013). Peran metode ekstraktive sangat penting karena metode ini berguna untuk memilih kalimat-kalimat penting dan menghindari redudansi sehingga dapat mencangkup banyak konsep.

Metode alternatif yang mampu memberikan good coverage pada ringkasan adalah clustering kalimat (Ge, dkk, 2011). Dalam mencapai good average pada ringkasan tidak terlepas dari koherensi cluster yang baik, yaitu menggunakan metode Similarity based Histogram Clustering (Sarkar, 2009). Metode Similarity based Histogram Clustering terbukti lebih baik jika dibandingkan dengan *Hierarchical Agglomerative Clustering* (HAC), *Single-Pass Clustering* dan *K-Nearest Neighbor Clustering*. Cluster-cluster yang sudah terbentuk selanjutnya akan diurutkan berdasarkan bobot kalimat yang paling penting dan dilakukan pemilihan kalimat representatif pada setiap cluster berdasarkan bobot kalimat menggunakan fungsi kombinasi antara fitur *local importance* dan fitur *global importance.* Selain itu dengan adanya penysuaian parameter histogram ratio pada Similarity besed Histogram Clustering dapat membantu meningkatkan performa dari sistem perangkuman.

Kalimat representatif tersebut adalah kalimat penting yang menjadi kalimat penyusun ringkasan dan merupakan salah satu penyelesaian untuk masalah saliency. Beberapa syarat yang harus dipenuhi, antara lain: kalimat yang dipilih harus mampu mewakili topik dari suatu cluster tertentu (Sarkar, 2009). Kalimat penting penyusun ringkasan harus mengandung informasi sebanyak mungkin dari dokumen sumber (He, dkk, 2008), kalimat menjadi penting jika ada keywords (kata kunci) yang mampu mempresentasikan dokumen. Pemilihan kalimat representatif dengan skema tersebut diharapkan mampu mewakili topik *cluster* secara lebih baik dan mampu meningkatkan kualitas hasil ringkasan.

Oleh karena itu, pada penelitian ini diusulkan sebuah fitur baru yaitu kata kunci *cluster* kalimat yang dikombinasikan dengan fitur *sentence information density* sebagai suatu strategi baru untuk memilih kalimat penting pada perangkuman multi-dokumen berdasarkan metode *clustering* kalimat dan pembobotan.

# Tujuan Penelitian

Tujuan penelitian yang dilakukan adalah untuk memperbaiki algoritma automatic summarization pada multi-dokumen. Langkah-langkah yang dilakukan dalam automatic summarization antara lain: preprocessing, cluster kalimat, pengurutan cluster, pemilihan kalimat representatif, dan penyusunan ringkasan. Perbaikannya terletak pada metode cluster kalimat dan pemilihan kalimat representatifnya dengan melakukan perhitungan score berdasarkan weighted graph model. Pada bagian pemilihan kalimat representatif akan diberikan ranking pada kalimat dan clustering multi-dokumen sehingga mendapatkan hasil perangkuman yang lebih optimal.

# Hipotesis

Hipotesa yang akan digunakan dalam penelitian ini adalah diharapkan hasilnya:

Perangkuman multi-dokumen yang dilakukan dengan metode cluster kalimat dan ranking pada kalimat diharapkan mendapatkan nilai ROUGE-1 berkisar 0.39 dan hasil ringkasan yang didapatkan lebih baik dari penelitian sebelumnya yaitu: ROUGE 0.37 (Sarkar, 2009).

# Tinjauan Pustaka

Dalam melakukan penelitian, terdapat beberapa tinjauan pustaka yang digunakan untuk membantu menyelesaikan permasalahan-permasalahan yang muncul dalam pengerjaan tesis ini. Tinjauan pustaka tersebut dapat dikaji sebagai berikut:

## Sentence Clustering-based Summarization of Multiple Text Document

Seiring dengan berkembangnya informasi pada dunia World Wide Web, informasi yang ada sangat banyak bahkan berlebihan dan memunculkan masalah bagi sebagian besar orang. Automatic Multi-document summarization dapat menjadi solusi yang berguna untuk mengatasi masalah kelebihan informasi yang ada di web. Segala jenis fitur dari perangkuman dapat membantu user untuk melihat hubungan antar dokumen dan cara mengatur informasi tersebut. Clustering menjadi salah satu penyelesaian perangkuman multi-dokumen dan dapat diterapkan pada web dimana bisa bergantung atau tidak pada bahasa yang digunakan.

Ada 3 faktor dari clustering yang sangat penting untuk melakukan multi- dokumen summarization, yaitu: cluster kalimat, pengurutan cluster, pemilihan kalimat representatif dari hasil cluster. Permasalahan yang dipelajari pada algortima ini bertujuan untuk mendapatkan algoritma cluster kalimat yang sukses dengan menggunakan sistem multi-dokumen summarization. Selain itu, banyak variasi yang dapat dipelajari untuk mendapatkan hasil sistem multi-dokumen yang terbaik. Kontribusi yang diberikan dari paper ini adakah mengenalkan mengenai cluster importance yang berarti sebuah metode yang diajukan untuk mengatasi kelemahan metode pengurutan cluster sederhana yang hanya bertumpu pada ukuran cluster. Fitur yang ditawarkan dari Sarkar adalah fitur local importance dan global importance.

Evaluasi dari perangkuman memanfaatkan perhitungan ROUGE. Dataset yang digunakan dalam penelitian ini menggunakan data set yang di dapat dari *Document Understanding Conference* (DUC) 2004. Datasetnya berupa 50 dokumen set, dimana masing-masing dokumen terdiri dari 10 berita berbahasa Inggris dan terdapat hasil perangkuman manual sebagai alat untuk mengavaluasi. Hasil perangkuman yang dibuat oleh pakar dihasilkan sekitar 665 byte atau kurang lebih 100 kata. Langkah-langkah pengerjaan dari paper ini diterapkan dalam penelitian yang akan dibuat. Langkah-langkahnya dapat dilihat pada bagian ruang lingkup yang sudah ditulis dalam penelitian ini.

Pada paper ini dilakukan uji coba yang lebih ditekankan pada pengurutan cluster dan pemilihan kalimat representatif. Metode yang diterapkan pada pengurutan cluster yaitu: pengurutan cluster dengan menghitung kalimat yang ada di cluster (*Cluster Ordering-SC*) dan pengurutan cluster yang berdasarkan informasi terbanyak (*ClusterOrdering-IR*). Metode yang digunakan untuk pemilihan representatif pada cluster yaitu: memilih secara random dari hasil cluster representatif tersebut (*Rep-Random*), memilih kandidat yang paling panjang dari cluster representatif (*Rep-Longest*), memilih kalimat yang kedekatan centroidnya paling kecil (*Rep-Centroid*), dan memilih kalimat dari cluster berdasarkan local dan global impotancenya (*Rep-LG*). Hasilnya berdasarkan

evaluasi ROUGE-1 dengan 95% confidance interval. Hasil yang dicapai pada paper dapat dilihat pada tabel 1.

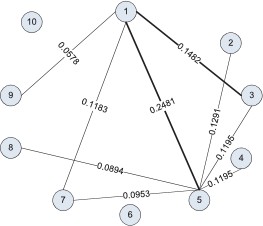
## Tabel 1

**Hasil ROUGE-1 score**

|  |  |
| --- | --- |
|  | ROUGE-1 score |
| Clustering Kalimat + ClusterOrdering-IR + Rep-LG | 0.3756 |
| Clustering Kalimat + ClusterOrdering-IR + Rep-Centroid | 0.3698 |
| Clustering Kalimat + ClusterOrdering-IR + Rep-Random | 0.3684 |
| Clustering Kalimat + ClusterOrdering-IR + Rep-Longest | 0.3678 |
| Clustering Kalimat + ClusterOrdering-SC + Rep-Centroid | 0.3545 |
| Clustering Kalimat + ClusterOrdering-SC + Rep-LG | 0.3541 |

1. **Weighted Graph Model Based Sentence Clustering and Ranking for Document Summarization**

Dalam peringkasan sebuah dokumen, algorima yang umum digunakan adalah algoritma yang merangkum sebuah dokumen tanpa menambahkan informasi apapun. Jenis algoritma yang lain adalah melakukan perangkuman dari dokumen berdasarkan topik atau query yang diberikan. Proses ektraksi sebuah kalimat yang penting dari sebuah dokumen dapat dihasilkan dari kalkulasi semua kalimat, dan memilih kalimat dengan skor yang tinggi. Membuat sebuah graph model untuk sebuah dokumen dan mengimplementasikan graph ranking algorithm seperti algoritma PageRank untuk menghitung skor dari sebuah kalimat adalah metode yang sangat efisien.



## Gambar 1 Ilustrasi Graph Model (Ge, 2011)

Di dalam graph model seperti pada Gambar 1, vertex merepresentasikan kalimat di dalam dokumen, dan bobot dari edge merepresentasikan kemiripan antar kalimat. Salah satu hal yang dapat di ekstrak dari sebuah kalimat adalah posisi kalimat, panjang kalimat, kata kunci kalimat, dan cue phrases dari setiap kalimat. Rangking dari kalimat didasarkan pada bobot-bobot fitur

tersebut. Mapping dari dokumen dibentuk dalam bentuk matrix dan algoritma faktorisasi matrix seperti *Singular Value Decomposition* (SVD), *Non- Negative Matrix Factorization* (NMF) dan *Symmetric Matrix Factorization* juga digunakan untuk proses perangkuman dokumen. Dari berbagai kalimat dibagi menjadi beberapa group dan setiap group akan dilakukan ektraksi kalimat pentingnya. Dari pandangan terkait dengan artinya, metode cluster kalimat akan menemukan sematiknya dengan cara menelusuri kumpulan sub topic yang tersebunyi (latent) dari dokumen yang secara tidak langsung memberikan informasi tambahan untuk cluster tersebut.

Dalam paper ini, diusulkan sebuah metode untuk melakukan perangkuman dokumen yang tidak menambahkan informasi tambahan, metode yang diusulkan adalah memanfaatkan informasi dari efek saling kalimat *(sentence mutual effects*) dan cluster kalimat. Penelitian dibentuk pemodelan graph dari dokumen yang ada, yang dilakukan perangkingan menggunakan graph- rangking algorithm. Sebuah matriks tambahan juga dibentuk dari kalimat yang diklasifikasikan berdasarkan perbedaan group menggunakan algoritma yang didasarkan pada *Sparse Non-negative Matrix Factorization* (SNMF). Kalimat yang berkaitan dalam satu cluster dan cluster tersebut memiliki rangking yang tinggi akan dipilih sebagai kalimat dalam menyusun rangkuman. Kontribusi pada penelitian ini adalah:

* 1. Algoritma perangkuman dokumen yang mengkombinasikan perangkingan kalimat dan clustering
  2. Sebuah metode clustering untuk kalimat berdasarkan SNMF
  3. Sebuah model graph berbobot (weighted graph) yang mempertimbangkan hubungan antar kalimat pada cluster dan pengurutan kalimat dalam suatu dokumen.

Dalam algoritma yang diusulkan pada penelitian ini jika dilakukan test pada multi dokumen, yang harus dilakukan adalah menggabungkan seluruh dokumen yang akan di rangkum kedalam satu dokumen besar setelah digabungkan menjadi satu dokumen dilakukan perangkingan setiap kalimat dari dokumen besar tadi.

## A New Feature-Fusion Sentence Selecting Strategy for Query-Focused Multi Document Summarization

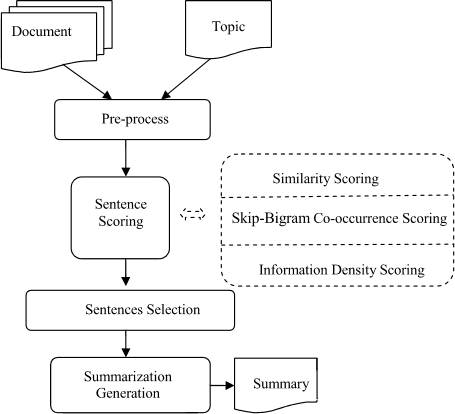
Langkah yang paling penting untuk menentukan fokus dari query perangkuman secara extractive dengan menentukan kalimat mana yang dapat dimasukkan ke dalam hasil final dari perangkuman. Pada paper ini, dijelasakan cara untuk megatur strategi pemilihan kalimat, mengidentifikasi kalimat dengan menggunakan high query relevance dan information density yang paling tinggi. Cara perhitungan nilai dari kalimat yang ada berdasarkan kemiripannya dan *Skip-Bigram co-occurrence* dengan query.

Ada dua fitur yang dapat menjadi pertimbangan hubungan antar query berdasarkan isi dan struktur query yang paling baik. Kemudian dilakukan perhitungan ulang kalimatnya menggunakan information density dengan fitur pembobotan dari teks graph yang berguna untuk menentukan posisi dari informasinya. Pada paper ini mengadopsi MMR (*Maximal Marginal*

*Relevance*) untuk mengurangi redudansi dan digunakan pada kalimat ekstraktive.

Sistem Arsitektur dari paper ini didesain menggunakan framework perangkuman ekstraktive dapat dilihat pada Gambar 2. Setelah proses pre- proses sudah dilakukan, kalimat penting dengan nilai tertinggi dan hasil ekstrak redudansinya yang paling sedikit akan masuk menjadi rangkuman akhir. Perhitungan penentuan nilai dari kalimat dapat dilihat pada bagian ruang lingkup.

Cara mendapatkan hasil evaluasi pada paper ini memanfaatkan perhitungan ROUGE dan diterapkan pada dataset yang berasal dari Document Understanding Conference (DUC) 2005. Isi dari dataset ini adalah 50 dokumen yang sudah dicluster dan setiap clusternya berisi 25-50 dokumen dan querynya. Masing-masing query terdiri dari judul dan beberapa gambaran mengenai querynya. Gambaran query yang dimaksudkan dianggap sebagai query yang dihasilkan oleh user dan 25-50 dokumen tersebut merupakan hasil retrival dari N dokumen. Untuk masing-masing cluster, dilakukan perangkuman secara manual.



## Gambar 2 Sistem Arsitektur (He, dkk, 2008)

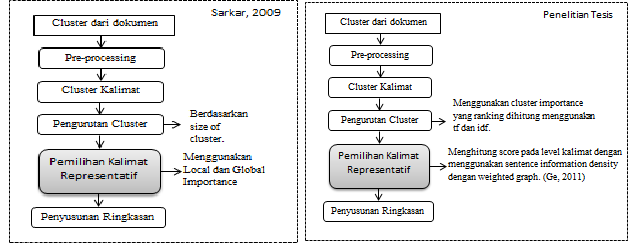
1. **Ruang Lingkup/Batasan Masalah**

Ruang lingkup yang akan digunakan dalam penelitian ini akan dijabarkan sebagai berikut:

## Rancangan Metode yang akan Digunakan

Proses yang digunakan dalam paper Sarkar (2009) dan penggabungan dengan metode weighted grapg (Ge, 2011), akan diterapkan pada penelitian Tesis yang akan dibuat. Rancangan Metode yang digunakan pada penelitian ini memiliki langkah-langkah sebagai berikut:

* + - Preprocessing
    - Cluster Kalimat
    - Pengurutan Cluster
    - Pemilihan Kalimat Representatif
    - Penyusunan Ringkasan



## Gambar 3 Blok Diagram Proses

Blok diagram proses dari penelitian yang digunakan dalam penelitian ini dapat dilihat pada Gambar 3. Penjelasan mengenai tahapan-tahapan dari blok diagram di atas:

1. Preprocessing

Pada tahapan preprocessing meliputi stop word dan menghilangkan tanda baca kecuali tanda titik (sebagai pembatas akhir kalimat). Stop words merupakan kata umum yang dimanfaatkan dalam information retrival yang biasanya muncul dalam jumlah besar dan dianggap tidak memiliki makna, contoh stop words dalam bahasa inggis: “of”, “the”, “and”, dan lain-lain.

1. Cluster Kali0mat

Cluster kalimat merupakan suatu komponen penting yang berbasis pada sistem perangkuman karena sub topik atau beberapa tema pada set dokumen harus dapat diidentifikasi menemukan persamaan dan perbedaan di dokumennya.

Perhitungan Cluster Kalimat:

* + *Similarity Measure*

*Cosine similarity* digunakan untuk menghitung kemiripan antara dua item dari sudut cosinus yang ada pada kedua item tersebut. Jika nilai cosinusnya adalah maka item tersebut benar-benar identik sedangkan jika nilainya 0 maka kedua item tersebut sama sekali tidak mempunyai kemiripan.

Rumus dari cosinus similarity:

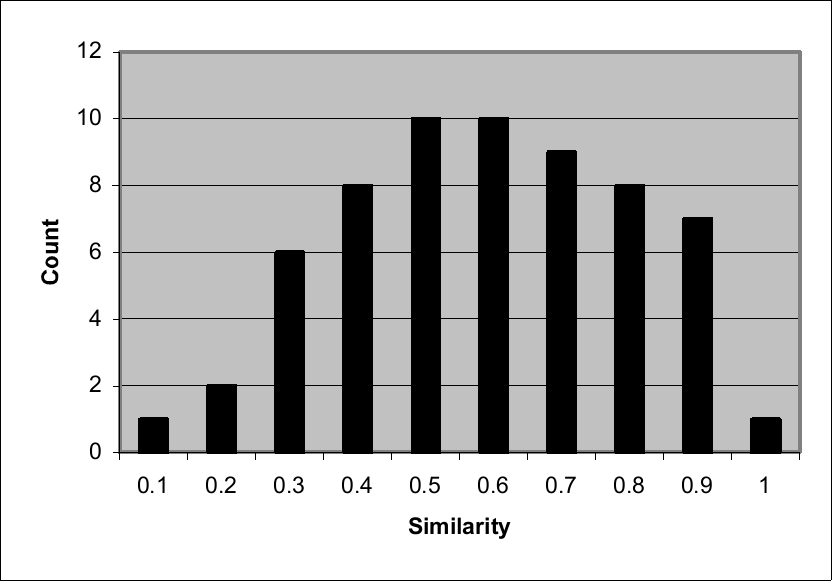
Sim (Si , Sj)=(2\*| Si ∩ Sj |) / ( |Si |+| Sj |)… (1)

Dimana Si dan Sj menunjukkan bahwa dua kalimat yang menunjukkan inputan. Perhitungan | Si ∩ Sj | menunjukkan jumlah kata antara dua kalimat yang sesuai dan | Si | menunjukkan panjang dari kalimat i-th dimana panjang dari kalimat = jumlah kata yang ada pada kalimat. Nilai similarity (ST) ini digunakan untuk menentukan dibentuknya bin baru (pada tahap similarity based histogram clustering) atau bukan.

* + Similarity based histogram clustering

Konsep utama dari model clustering dengan menggunakan *Similarity based Histogram Clustering* (SHC) adalah menjaga setiap *cluster*-*cluster* yang terbentuk tetap koheren/terpadu/berkaitan kapan saja (Hammouda dan Kamel, 2003). Dalam Similarity based Histogram Clustering diperkenalkan konsep *cluster similarity histogram*. Konsep tersebut adalah konsep digunakan untuk merepresentasikan koherensi dari *cluster*.

Similarity based Histogram Clustering adalah representasi statistik dari suatu distribusi kesamaan pasangan antar anggota yang ada pada suatu cluster. Representasinya menunjukkan jumlah dari *bin* dalam *histogram* artinya interval nilai *similarity* tertentu. Pada masing- masing *bin* terdiri dari jumlah *similarity* dari masing-masing pasangan anggota *cluster* yang memenuhi interval pada *bin*.



**Gambar 4 Distribusi *Similarity* Setiap Pasangan Anggota *Cluster***

## (Sarkar, 2009)

Cara yang terbaik untuk mendapatkan nilai dari derajat koherensinya yang tinggi dalam cluster, yaitu dengan mempertahankan derajat similarity antar anggota tetap tinggi. Dalam similarity histogram, berarti menjaga distribusi *similarity* agar cenderung ke kanan (ke arah nilai 1 yaitu nilai *similarity* terbesar). Grafik lebih lanjut dapat dilihat pada Gambar 4.

Pada grafik Gambar 4 terdapat 10 (sepuluh) *bin*. Masing-masing tinggi bin merepresentasikan total jumlah dari nilai pasangan similarity sesuai dengan interval dari bin tersebut. Koherensi antar anggota cluster yang ingin dicapai adalah hal utama yang dioptimalkan pada Similarity based Histogram Clustering. Dibutuhkan sebuah nilai *threshold* yang dimanfaatkan untuk menentukan tingkat toleransi sebuah *cluster* dalam menerima anggota cluster baru.

Setiap kalimat yang menjadi anggota suatu *cluster* tertentu ditentukan berdasarkan *threshold* dan kondisi dari *histogram ratio*. Jika memenuhi kriteria suatu kalimat dapat masuk ke dalam suatu *cluster* tertentu. Sedangkan jika suatu kalimat tidak memenuhi kriteria dari setiap *cluster* yang telah terbentuk sebelumnya maka kalimat tersebut harus membentuk suatu *cluster* baru.

Setiap cluster memiliki nilai histogram rasio, Setiap kali ada kalimat yang digabungkan ke dalam sebuah cluster akan menyebabkan perubahan rasio histogram yang disebut HRnew. Nilai histogram rasio sebelum ditambahkan kalimat disebut HRold. Parameter *ɛ* (*epsilon*) adalah sebuah *threshold* sebagai ambang batas selesih antara *HRold* dengan *HRnew*. Pada proses pembentukan *cluster*, nilai (*HRold*) dan nilai (*HRnew*) dibandingkan. Jika nilai *HRnew* lebih atau sama dengan nilai *HRold*, maka kalimat tersebut ditambahkan ke dalam suatu *cluster*. Jika nilai *HRnew* lebih rendah dari nilai *HRold* namun nilai *HRnew* masih berada diatas *HRmin* dan selisih antara *HRold* dengan *HRnew* tidak lebih dari *ɛ* maka kalimat juga ditambahkan sedangkan selain itu kalimat tidak ditambahkan. *Clustering* kalimat dengan metode SHC dapat menjamin setiap *cluster* kalimat yang terbentuk setidaknya memiliki sebuah anggota oleh karena itu proses pembentukan *cluster* kalimat terus dilakukan terhadap semua kalimat *input*.

1. Pengurutan Cluster

Proses pengurutan cluster dilakukan dengan algoritma unsupervised sehingga jumlah cluster yang akan terbentuk tidak dapat ditentukan. Penentuan jumlah cluster ini merupakan hal yang sangat crusial dikarenakan jumlah cluster yang terbentuk akan menentukan mana kalimat representatif yang akan dirangkum. Metode yang paling mudah untuk mengurutkan cluster adalah dengan menghitung jumlah kalimat yang ada di dalam cluster dengan asumsi sebuah cluster akan jadi penting jika terdapat banyak kalimat. Jika menggunakan metode sederhana tersebut maka akan muncul masalah pada saat terdapat beberapa cluster yang memiliki jumlah kalimat yang sama, atau

sebuah cluster terdapat banyak kalimat-kalimat pendek yang tidak informatif. Dimana kalimat-kalimat pendek ini hanya menambah jumlah kalimat tanpa menambahkan informasi.

*Cluster importance* adalah sebuah metode yang diajukan untuk mengatasi kelemahan metode pengurutan *cluster* sederhana yang hanya bertumpu pada ukuran *cluster* (Sarkar, 2009). Konsep dari *cluster importance* adalah mengurutkan berdasarkan nilai penjumlahan bobot dari kata-kata yang merupakan kata penting yang terkandung dalam *cluster*. Kata penting yang dimaksud adalah kata-kata yang sering muncul (*frequent*) pada semua dokumen *input*. Sebuah *threshold* (*θ*) ditetapkan untuk menentukan apakah suatu kata tersebut termasuk kata penting atau tidak terhadap seluruh dokumen *input*. Jika frekuensi suatu kata memenuhi *threshold θ* maka kata tersebut dianggap sebagai kata penting. Pendekatan *cluster importance* bertujuan mengukur pentingnya suatu *cluster* berdasarkan jumlah kata-kata penting yang ada pada suatu *cluster*. Tingkat kepentingan sebuah cluster yang dihitung dengan rumus:

*Weight*(*cj* )  log(1  *count*(*w*))|)… (2)

*w**cj*

Dimana bobot dari *cluster c* ke-*j* dinotasikan dengan *Weight(cj)*, *count (w)* adalah jumlah dari kata *w* pada koleksi *input* dan *count (w)*. Pengurutan cluster dilakukan berdasarkan nilai *Weight(cj)* tersebut*.* Semakin besar nilai *Weight(cj)* menunjukkan cluster tersebut lebih penting. Cluster yang lebih penting berarti memiliki frekuensi kata penting yang banyak.

Sebuah kalimat dari sebuah cluster akan menjadi kalimat penyusun ringkasan akhir. Proses tersebut dilakukan secara menerus sesuai dengan panjang ringkasan yang diharapkan sehingga cluster dengan nilai bobot x tertinggi akan terpilih sebagai cluster kandidat ringkasan. Proses pemilihan kalimat representatif dari sebuah cluster dilakukan pada tahap pemilihan kalimat representatif.

1. Pemilihan Kalimat Representatif

Fase ini merupakan kontribusi dan fokus utama dari penelitian penulis. Kontribusi yang didapatkan adalah penggantian algoritma pada bagian kalimat representatifnya pada paper Sarkar,2009 menjadi algoritma weighted graph (Ge, 2011). Jadi kalimat yang dipilih hanya berupa hasil random, jarak antar kata, pemilihan kalimat berdasarkan kemiripan kalimat, dan pemilihan berdasarkan local dan global importance. Pemilihan kalimat representatif dilakukan dengan cara menghitung nilai score pada level kalimat. Score setiap kalimat dihitung berdasarkan penelitian He (2008), yaitu menghitung *sentence information density* yang mengacu pada posisional sentence graph dalam suatu cluster kalimat. Nilai similarity untuk setiap kalimat yang ada pada cluster dihitung dengan kalimat lainnya. Jika nilai similarity kedua kalimat lebih dari threshold (α) maka antar kelimat tersebut dianggap beralasi

yang akan membentuk sebuah edge. Kemudian dibentuk similatiry matrix yang digunakan untuk membentuk graph kalimat yang merepresentasikan position information. Graph digambarkan sebagai P=(V,E), dimana P merepresentasikan graph, V={s1, s2,…,sn} adalah vertex pada graph yang merepresentasikan kalimat-kalimat dalam suatu cluster, dan E={(si, sj)} adalah kumpulan dari suatu edge pada graph dimana bobot edge tersebut dihitung berdasarkan similarity antar dua kalimat dalam cluster.

Graph P dibangun berdasarkan kalimat-kalimat dalam cluster. Saat pertama graph P kosong, setelah itu semua kalimat dalam suatu cluster dimasukan sebagai *vertex*. Langkah kedua hitung nilai similarity untuk setiap pasangan kalimat dalam P, jika nilai similarity suatu pasangan kalimat memenuhi threshold α maka edge dibentuk dan bobot pasangan kalimat tersebut adalah nilai similarity yang dimilikinya. Ketika graph telah dibangun, fitur sentence information density dihitung dengan rumus:

*F* (*s* ) 

*Wskj*,

*sid kj*

max *W*

*l*{1,2,..*n*} *lj*

*s*

……………….……………………….………………(3)

Dimana jumlah kalimat s pada cluster ke-j ditunjukkan dengan n, *Wskj* adalah penjumlahan bobot dari semua edge yang datang dari kalimat s ke-k pada

max *W*

*s*

cluster ke-j, sedangkan *l*{1,2,..*n*} *lj*

adalah bobot edge maksimum diantara

semua kalimat yang ada pada cluster ke-j.

1. Penyusunan Ringkasan

Setelah mengelompokkan kalimat, cluster diperintahkan menggunakan algoritma pengurutan cluster. Satu kalimat representatif dari setiap cluster dipilih dengan memanfaatkan algoritma representatif pemilihan. Dipilih satu kalimat yang sudah tercluster pada posisi paling atas dan kemudian memilih memilih kalimat dari subsequence cluster berada pada daftar cluster urutan sampai panjang dari ringkasan ditemukan.

## Data yang Digunakan

Dataset yang digunakan dalam penelitian ini menggunakan data set yang di dapat dari Document Understanding Conference (DUC) 2004. Datasetnya berupa 50 dokumen set, dimana masing-masing dokumen terdiri dari 10 berita berbahasa Inggris dan terdapat hasil perangkuman manual sebagai alat untuk mengavaluasi. Hasil perangkuman yang dibuat oleh pakar dihasilkan sekitar 665 byte atau kurang lebih 100 kata.

## Target Pencapaian Penelitian

Target penelitian yang ingin dicapai adalah menggabungkan kedua metode teks summarization yang berdasarkan sentence clustering pada multi-dokumen.

Hasil penelitian diharapkan mendapatkan nilai ROUGE-1 berkisar 0.39 dan hasil ringkasan yang didapatkan lebih baik dari penelitian sebelumnya yaitu mencapai nilai ROUGE 0.37 (Sarkar, 2009).

## Batasan dari Penelitian Tesis

Batasan dari tesis yang dibuat dalam penelitian ini adalah:

* + - Permasalahan yang diteliti adalah summarzation berdasarkan kategori ekstraktive bukan abstractive.
    - Tidak menentukan urutan di dalam kalimat.
    - Dataset yang digunakan adalah dataset DUC 2004 yang berisi dokumen berita berbahasa Inggris.
    - Scope penelitian tidak berada pada text pre-processing sehingga tahap text pre-processing menggunakan library yang sudah ada.
    - Metode evaluasi yang digunakan mengacu pada Lin (2004) yang menguji kualitas algoritma (ROUGE). ROUGE kepanjangan dari Recall Outstanding Understudy of Grading Evaluation.

## Uji Coba

Uji coba dilakukan pada 50 himpunan dokumen dan masing-masing himpunan dokumen memiliki 10 dokumen berita berbahasa inggris. Dari 50 himpunan yang ada, dibagi menjadi 2 bagian. Dua puluh lima himpunan dokumen digunakan untuk menguji parameter-parameter yang digunakan pada metode penelitian ini. Tujuan pengujian parameter diharapkan dapat menghasilkan nilai ROUGE yang optimal. Setelah parameter-parameter ditemukan, dilakukan uji coba terhadap dua puluh lima himpunan dokumen yang lain. Parameter-parameter yang akan diestimasi pada penelitian ini dapat dilihat pada tabel 2.

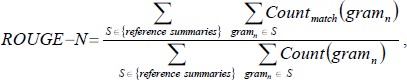
## Tabel 2 Parameter Uji Coba

|  |  |
| --- | --- |
| **Notasi** | **Keterangan** |
| *HRmin* | Batas nilai minimum dari *Histogram Ratio* |
| *ɛ* | Batas selesih maksimum antara *HRold* dengan *HRnew* |
| *ST* | Batas *similarity bin* pada perhitungan *histogram ratio* |
| *θ* | Batas frekuensi minimal kata *w* dalam proses pengurutan *cluster* |
| *α* | Nilai *threshold* untuk menentukan pembentukkan *edge* antar kalimat pada fitur *sentence information density* |

* 1. **Evaluasi**

ROUGE merupakan suatu wadah yang digunakan untuk mengevaluasi hasil rangkuman. Evaluasi yang digunakan dalam penelitian ini memanfaatkan ROUGE N-gram (Lin, 2004) karena perhitungan hasil evaluasi untuk text summarization cocok menggunakan ROUGE (Recall Outstanding Understudy of

Grading Evaluation). Pada paper yang dijadiakan acuan milik Sarkar, He, dan Ge memanfaatkan ROUGE. Cara kerja ROUGE adalah dengan hasil perhitungan human summarization digabung dengan auto summarization. Parameter yang menjadi perhitungan dalam ROUGE meliputi N-gram, urutan kata, pasangan kata antara hasil rangkuman yang dihasilkan dari ringkasan komputer yang akan dievaluasi dengan hasil rangkuman yang dibuat oleh manusia.

Rumus yang digunakan pada perhitungan ROUGE:

………………………(5)

Dimana N merupakan panjang dari N-gram, Countmatch(N-gram) menunjukkan jumlah maksimum dari N-gram yang muncul pada ringkasan kandidat dan ringkasan sebagai referensi. Count(N-gram) adalah jumlah dari N-gram pada rangkuman yang dijadikan referensi. Pada paper ini fungsi ROUGE-N yang digunakan adalah ROUGE dengan nilai N = 1. Nilai akhir dari ROUGE-N akhir adalah nilai ROUGE-N terbesar yang dihasilkan dari pasangan rangkuman hasil sistem dan rangkuman referensi. Nilai ROUGE-N dihitung pada setiap pasangan ringkasan kandidat sc dan ringkasan referensi rsi. Perhitungan ROUGE-N tersebut diadopsi dari Lin (2004) yang ditunjukkan pada rumus:

ROUGE-Nmulti =argmaxi ROUGE-N(sc,rsi) (6)

# Jadwal Penelitian

Jadwal dari masing-masing tahap untuk penelitian ini dapat dilihat pada

tabel 3.

## Tabel 3 Jadwal Penelitian

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| No | Kegiatan | Bulan | | | | |
| 1 | 2 | 3 | 4 | 5 |
| 1 | Studi  Literatur |  |  |  |  |  |
| 2 | Pengumpulan  Data |  |  |  |  |  |
| 3 | Perancangan  Struktur Data |  |  |  |  |  |
| 4 | Implementasi |  |  |  |  |  |
| 5 | Uji Coba |  |  |  |  |  |
| 6 | Penulisan  Laporan |  |  |  |  |  |

1. **Daftar Pustaka**
2. Ge, S. S., Zhang Z., dan He, H. (2011), “Weighted Graph Model Based Sentence Clustering and Ranking for Document Summarization” *Proceeding of 2011 4th International Conference on Interaction Sciences (ICIS)*, National University of Singapore, Singapore, hal. 90-95.
3. Gupta, V. (2010), “A Survey of Text Summarization Extractive Tecniquies”, *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, hal. 258-268.
4. Hammouda, K. M. dan Kamel, M. S. (2003), “Incremental Document Clustering Using Cluster Similarity Histograms” *Proceeding of the 2003 IEEE/WIC International Conference on Web Intelligence*, Eds: Liu, J. et al., University of Waterloo, Halifax, Canada, hal. 597-601.
5. He, T., Li F., Shao, W., Chen, J., dan Ma, L. (2008), “A New Feature- Fusion Sentence Selecting Strategy for Query-Focused Multi-document Summarization”, *Proceeding of International Conference Advance Language Processing and Web Information Technology*, Eds: Ock C. et al., University of Normal, Wuhan, China, hal. 81-86.
6. Hovy, E. H. and C.-Y. Lin. 2001, “Automating text summarization in SUMMARIST”, In Mani and Maybury (2001), 81-97.
7. Lin, C. Y. (2004), “ROUGE: a Package for Automatic Evaluation of Summaries”, *In Proceedings of Workshop on Text Summarization Brances Out*, Eds: Moens, M. F. dan Szpakowicz, S., Association for Computational Linguistics, Barcelona, hal. 74-81.
8. Ouyang, Y., Li W., Zhang R., Li S., dan Lu Q. (2013), “A Progressive Sentence Selection Strategy for Document Summarization”, *Journal of information Precessing and Management*. Vol. 49, Issue 1, hal. 213-221.
9. Sarkar, K. (2009), “Sentence Clustering-based Summarization of Multiple Text Documents”, *International Journal of Computing Science and Communication Technologies*, Vol. 2, No. 1, hal. 325-335.
10. Wan, X. dan Yang, J. (2008), “Multi-Document Summarization Using Cluster-Based Link Analysis”, *Proceedings of the 31st annual international ACM SIGIR conference on Research and Development in Information Retrieval*, Eds: Chua T. S. et al., Association for Computational Linguistics, New York, hal. 181-184.