TEMU KEMBALI INFORMASI BERDASARKAN LOKASI PADA DOKUMEN YANG DIKELOMPOKKAN MENGGUNAKAN METODE *CENTROID LINKAGE HIERARCHICAL*

Nadia Damayanti[1], Nur Rosyid Mubtada'i, S.Kom, M.Kom[2], Afrida Helen S.T, M.Kom[2]
[1]Mahasiswa, [2]Dosen Pembimbing
Jurusan Teknik Informatika
Politeknik Elektronika Negeri Surabaya
Institut Teknologi Sepuluh Nopember Surabaya Kampus PENS-ITS Keputih Sukolilo Surabaya 6011, Indonesia
Tel: +62-85655616926, 0351-369234
Email: nadia.damayanti240989@gmail.co.id

Abstrak

Pencarian informasi berdasar kata kunci dapat membantu pengguna ketika ingin mengetahui informasi yang berhubungan dengan kata kunci yang dicari. Begitupun dengan pencarian informasi ketika pengguna ingin mengetahui kelompok dokumen yang memuat lokasi tertentu yang sama. Karenanya dibutuhkan suatu sistem yang memberikan informasi kepada pengguna yang ingin melakukan pengelompokkan dokumen berdasar lokasi tertentu yang sama. Dengan menerapkan temu kembali informasi dalam mencari kata kunci lokasi pada dokumen dan metode Centroid Linkage Hierarchical sebagai metode pengelompokkan data diharapkan dapat memberikan hasil yang optimal dalam mengelompokkan dokumen dan hanya akan mengambil informasi yang mempunyai tingkat kepentingan tinggi.

Kata kunci : *Clustering, Centroid Linkage Hierarchical, Temu Kembali Informasi*

1. PENDAHULUAN

1.1. LATAR BELAKANG

Kemajuan teknologi yang semakin pesat telah memaksa manusia untuk berusaha mengikutinya. Teknologi tersebut dapat digunakan oleh semua kalangan yang dapat memanfaatkannya untuk berbagai keperluan. Teknologi tersebut memudahkan mereka dalam memenuhi kebutuhan dengan lebih cepat, lebih efisien, dan tepat sehingga waktu yang akan dipergunakan dapat semakin dipangkas. Hal tersebut seiring pula dengan perkembangan Teknik Informatika yang muncul pada cabang ilmu baru yaitu Temu Kembali Informasi Informasi *(Information Retrieval).* Pencarian informasi (*Information Retrieval*) [1] adalah salah satu cabang ilmu yang bertujuan untuk membantu pengguna dalam menemukan informasi yang relevan dengan kebutuhan mereka dalam waktu singkat.

Pencarian informasi berdasarkan *keyword* sangat berguna untuk pencarian terarah dan membantu *user* ketika ingin mengetahui informasi yang berhubungan dengan *keyword* yang dicari. Begitu pula dengan pencarian informasi ketika pengguna ingin mengetahui kelompok dokumen yang memuat lokasi tertentu yang sama. Maka dari itu dibutuhkan suatu sistem yang memberikan informasi kepada pengguna yang ingin melakukan pengelompokkan dokumen berdasarkan lokasi tertentu yang sama.

Dalam proyek akhir ini akan dibahas tentang bagaimana cara untuk mendapatkan informasi lokasi dari dokumen yang menggunakan Temu Kembali Informasi dengan tahapan antara lain *Parsing, Filtering,* Analisa Semantik, Translasi dan bagaimana cara untuk mengelompokkan dokumen tersebut berdasarkan kata kunci lokasi yang telah ditemukan. Metode yang digunakan untuk proses pengelompokkan dalam proyek akhir ini adalah metode *Centroid Linkage Hierarchical Method*. Besarnya data pada masing-masing hasil pengklasteran selanjutnya akan digunakan untuk menentukan hasil pengelompokkan dokumen.

1.2. PERMASALAHAN

Adapun permasalahan yang ada pada system ini yaitu sebagai berikut:

1. Bagaimana penerapan Temu Kembali Informasi Teks dalam mengelompokkan dokumen berdasarkan kata kunci lokasi.
2. Bagaimana metode *Centroid Linkage Hierarchical* Method dapat dipergunakan

dalam membuat pengelompokkan dokumen berdasarkan kata kunci lokasi
3. Bagaimana membuat sebuah aplikasi dengan menggunakan metode *Centroid Linkage Hierarchical Method.*

1.3. BATASAN MASALAH

Batasan masalah dalam penerapan teknologi ini adalah:

1. Pengelompokkan pada proyek akhir ini hanya menggunakan satu macam kata kunci saja yaitu kata kunci lokasi.
2. Pengklasteran nama lokasi otomatis.
3. Data yang digunakan sebagai data input adalah *online.*
4. Jumlah kata kunci masukan paling banyak 2 kata.
5. Dokumen yang akan diolah oleh *cluster* diambil dari beberapa situs berita *online.*

1.4. TUJUAN

Proyek akhir yang berjudul "Temu Kembali Informasi Berdasarkan Lokasi Pada Dokumen Yang Dikelompokkan Menggunakan Metode *Centroid Linkage Hierarchical*" ini bertujuan antara lain untuk :

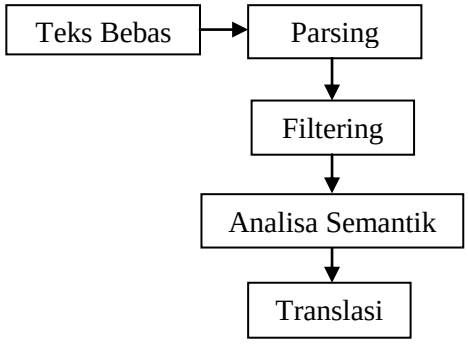
1. Mendapatkan pola kesamaan dalam dokumen berdasarkan lokasinya dengan menggunakan Temu Kembali Informasi *(Information Retrieval).*
2. Menerapkan algoritma *Centroid Linkage Hierarchical Method* dalam pengelompokkan dokumen berdasarkan kata kunci lokasi.

2. TEORI PENUNJANG

2.1. Temu Kembali Informasi

Ketika sebuah aplikasi membutuhkan fasilitas untuk kegiatan penyimpanan, penyediaan representasi, identifikasi, serta pencarian atau penelusuran dokumen yang sesuai pada suatu *database*, maka dibutuhkan sistem temu kembali informasi. Proses sistem temu kembali informasi yang dikembangkan dalam proyek akhir ini meliputi 4 proses utama yaitu *parsing*, *filtering*, analisa semantik, dan translasi.

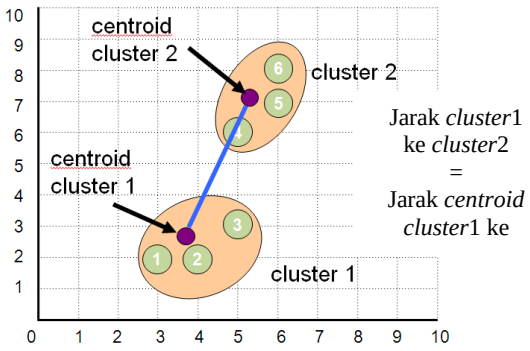
Berdasarkan makalah yang ditulis oleh Surya Sumpeno dkk[2], proses parsing berfungsi untuk memecah kata, proses filtering berfungsi untuk menyaring kata-kata yang tidak dibutuhkan dengan menggunakan algoritma stopword, proses analisa semantik digunakan untuk menganalisa kata-kata yang berhubungan atau biasanya mengawali kata lokasi, terakhir proses translasi didapatkan kata-kata lokasi yang dibutuhkan sebagai kata kunci. Berikut ini adalah blok diagram temu kembali informasi :



Gambar 2.1. Blok Diagram Temu Kembali Informasi

2.2. *Centroid Linkage Hierarchical Method*

Berdasarkan proyek akhir yang dikerjakan Hervilorra E. [3] dijelaskan bahwa *Centroid Linkage Hierarchical Method* adalah proses pengklasteran yang didasarkan pada jarak antar centroidnya. Metode ini baik untuk kasus *clustering* dengan *normal data set distribution*. Akan tetapi metode ini tidak cocok untuk data yang mengandung *outlier*. Ilustrasi dari algoritma *Centroid Linkage Hierarchical Method* digambarkan seperti berikut :



Gambar 2.2 Ilustrasi Algoritma *Centroid Linkage Hierarchical Method*

Sumber: Eldira, Hervilorra, Web Mining Untuk Pencarian Dokumen Bahasa Inggris Menggunakan Hill Climbing Automatic Cluster 2010, Politeknik Elektronika Negeri Surabaya,ITS.

Algoritma *Centroid Linkage Hierarchical Method* :

1. Diasumsikan setiap data dianggap sebagai *cluster*. Kalau *n*=jumlah data dan *c*=jumlah *cluster*, berarti ada *c*=*n*.
2. Menghitung jarak antar *cluster* dengan *Euclidian distance.*
3. Mencari 2 *cluster* yang mempunyai jarak *centroid* antar *cluster* yang paling minimal dan digabungkan *(merge)* kedalam *cluster* baru (sehingga c=c-1).
4. Kembali ke langkah 3, dan diulangi sampai dicapai *cluster* yang diinginkan.

2.3. Varian Suatu Cluster

Berdasarkan proyek akhir yang dikerjakan oleh Hervilorra E.[3], dijelaskan bahwa varian suatu *cluster* digunakan agar mendapatkan jumlah *cluster* yang tepat secara otomatis.

Dijelaskan[3] bahwa suatu teknik analisa *multivariate* (banyak variabel) untuk mencari dan mengorganisasi informasi tentang variabel tersebut sehingga secara relatif dapat dikelompokkan dalam bentuk yang homogen dalam sebuah *cluster* adalah analisa *cluster*. Secara umum, dapat dikatakan sebagai proses menganalisa baik tidaknya suatu proses pembentukan *cluster*. Analisa *cluster* dapat diperoleh dari kepadatan cluster yang dibentuk (*cluster density*). Kepadatan suatu *cluster* dapat ditentukan dengan variance within cluster (Vw) dan *variance between cluster* (Vb) dimana varian tiap tahap pembentukan cluster dapat dihitung dengan rumus:

$$Vc^2 = \frac{1}{n_c - 1}\sum_{i=1}^{n}\left(y_i - \overline{y}_c\right)^2 \qquad \text{.....(2.1)}$$

Dimana :
Vc^2 = varian pada *cluster c*
c = 1..k, dimana k = jumlah cluster
n_c = jumlah data pada *cluster c*
$\underline{y_i}$ = data ke-*i* pada suatu cluster
$\overline{y_i}$ = rata-rata dari data pada suatu *cluster*

Selanjutnya dari nilai varian di atas, kita dapat menghitung nilai variance within cluster (Vw) dengan rumus:

$$Vw = \frac{1}{N - c}\sum_{i=1}^{c}\left(n_i - 1\right)\cdot Vi^2 \qquad \text{....}\cdots\cdots\cdots(2.2)$$

Dimana, N = Jumlah semua data
n_i = Jumlah data *cluster* i
Vi = Varian pada *cluster* i

Dan nilai variance between cluster (Vb) dengan rumus:

$$Vb = \frac{1}{c - 1}\sum_{i=1}^{c} n_i\left(\overline{y}_i - \overline{y}\right)^2 \qquad \text{.....(2.3)}$$

Dimana, $\overline{y}$ = rata-rata dari $\overline{yi}$

Salah satu metode yang digunakan untuk menentukan cluster yang ideal adalah batasan variance, yaitu dengan menghitung kepadatan cluster berupa variance within cluster (Vw) dan variance between cluster (Vb) [5]. Cluster yang ideal mempunyai Vw minimum yang merepresentasikan internal homogenity dan maksimum Vb yang menyatakan external homogenity.

$$V = \frac{Vw}{Vb} \qquad \text{.....(2.4)}$$

2.4. Algoritma Crawling

Berdasarkan proyek akhir yang telah dikerjakan oleh M. Badrullami[4] bahwa agar suatu aplikasi dapat memiliki fasilitas dalam mengumpulkan informasi melalui jaringan internet *(online*) dan kemudian hasilnya akan disimpan dalam suatu storage maka dibutuhkan sebuah web *crawler*. Algortima *crawling* dalam proyek akhir ini mengacu pada algoritma *crawling* yang telah dikerjakan oleh M. Badrullami [4].

Gambar 2.3 Arsitektur Sistem *Web Crawler*

Sumber: Badrullami, Moh, *Rancang Bangun Aplikasi Server Crawling Berita Online Sebagai Penyedia Berita Up To Date Pada Handphone Yang Mendukung WAP*, 2010, Politeknik Elektronika Negeri Surabaya,ITS.

Berdasarkan gambar sebelumnya, *crawler* diawali dengan adanya daftar URL yang akan dikunjungi, bisa juga disebut dengan *seeds*. Setelah *crawler* mengunjungi URL tersebut, kemudian mengidentifikasi semua *hyperlink* dari halaman itu dan menambahkan kembali kedalam seeds. Hal ini dinamakan *crawl frontier*. Setelah web crawler mengunjungi halaman –halaman web yang ditentukan di dalam *seeds*, maka web *crawler* membawa data – data yang dicari oleh pengguna kemudian menyimpannya ke dalam *storage*.

2.5. Reguler Expression *(REGEX)*

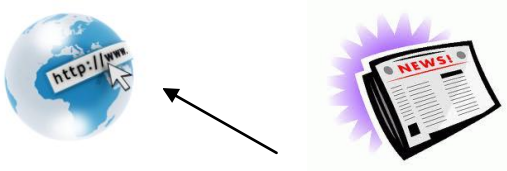
Mengacu pada proyek akhir yang dikerjakan oleh M.Badrullami [4] ketika sebuah aplikasi menggunakan dokumen sumber yang langsung di ambil dari situs berita *online* tentu saja format dokumen yag didapatkan akan memiliki pola-pola string tertentu dimana nantinya akan menyebabkan proses temu kembali informasi menjadi kesulitan dalam melakukan proses analisa. Maka *regex* digunakan untuk pencarian string dengan cara menentukan pattern string tersebut, sehingga string yang didapat dari proses *crawling* dapat di*filter*.

3. PERANCANGAN DAN PEMBUATAN SISTEM

3.1. Gambaran Umum

Urutan proses sistemnya adalah sebagai berikut:

- Menentukan beberapa *situs* berita *online* yang digunakan sebagai dokumen yang akan dicari kata kunci lokasinya.
  Dalam langkah ini, telah disiapkan beberapa situs penyedia berita *online* yang senantiasa selalu update.



http://www.tempointeraktif.com

internet berita

Gambar 3.1 Ilustrasi Pengambilan Dokumen Dari Internet Melalui Situs Tertentu

Sumber: www.google.com

- Regular expression
  merupakan sebuah pola dari suatu *string*. Regex digunakan untuk pencarian *string* dengan cara menentukan pola *string* tersebut yaitu dengan menghilangkan atau mengubah *string* dokumen dari internet untuk menjadikan format yang standart sehingga dapat diproses dalam temu kembali informasi.

- Temu kembali informasi
  Melakukan proses temu kembali informasi untuk dokumen sumber yang akan dilakukan proses pengklasteran. Proses temu kembali informasi itu sendiri terdiri dari *parsing*, *filtering*, analisa semantik, dan translasi.

- Menghitung jumlah kata kunci lokasi yang terdapat di dalam dokumen.
  Dalam proses ini adalah dilakukan penghitungan jumlah kata kunci lokasi yang ada di dalam setiap dokumen.

- Proses *clustering* dengan CLHM (*Centroid Linkage Hierarchical Method)*

Dilakukan proses klasterisasi sesuai dengan kata kunci lokasi yang dihasilkan pada tiap-tiap dokumen.

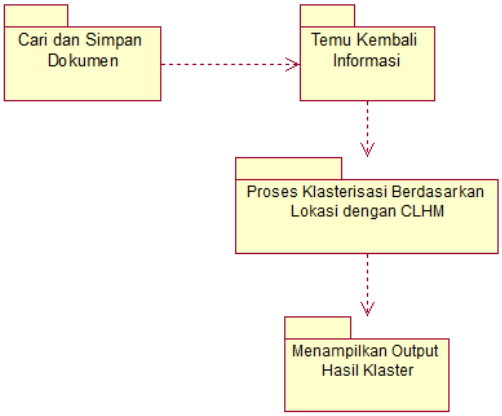
3.2 Desain Sistem

3.2.1. Desain Input

Input dari sistem proyek akhir ini adalah merupakan kata kunci yang dimasukkan oleh pengguna dan data yang diambil dari internet pada saat itu juga.

3.2.2. Desain Proses

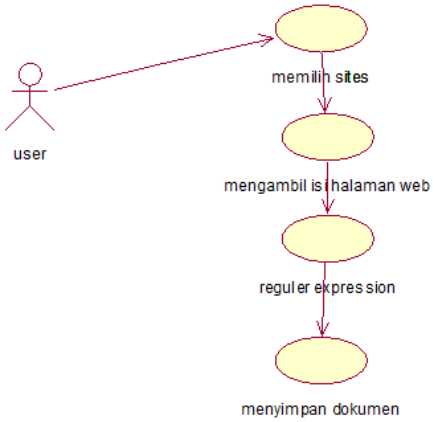
Use-Case Utama pada gambar 3.2 adalah gambaran sistem secara garis besar yang dibedakan menjadi empat proses utama, yaitu proses cari dan simpan dokumen online, proses temu kembali informasi, proses pengklasteran dengan algoritma *Centroid Linkage Hierarchical Method*, dan menampilkan hasil proses klasterisasi berdasarkan kata kunci lokasi terhadap dokumen yang ada. Untuk penjelasan lebih lanjut mengenai proses dan alur yang terjadi pada keempat sistem tersebut dijelaskan dengan menggunakan *Use-case diagram* dari tiap sistem.



Gambar 3.2 Use Case Utama

3.2.2.1. Use Case Diagram Pencarian dan Penyimpanan Dokumen *Online*

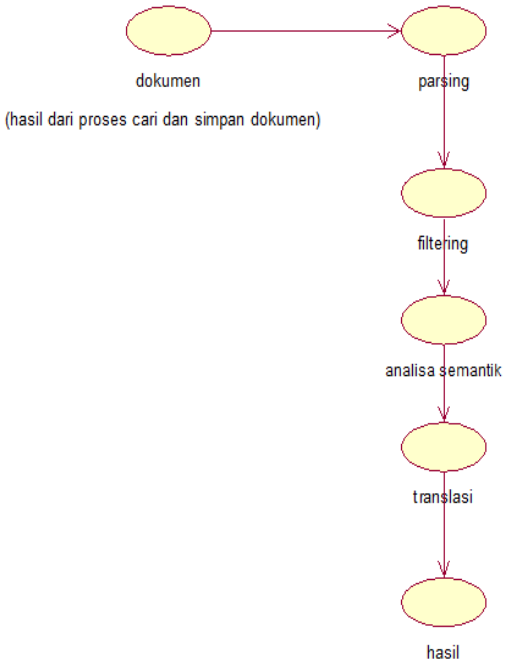
Gambar 3.3 adalah merupakan use-case diagram untuk proses pencarian dan penyimpanan dokumen yang diambil dari internet. Dimana situs sumber yang akan dijadikan sebagai situs tempat pencarian dokumen telah ditentukan sebelumnya. Pada proses ini terjadi pengecekan apakah situs yang dimasukkan berupa situs berita atau bukan.



Gambar 3.3 Use Case Diagram Proses Pencarian Dan Penyimpanan Dokumen Dari Internet

3.2.2.2. Use Case Diagram Temu Kembali Informasi

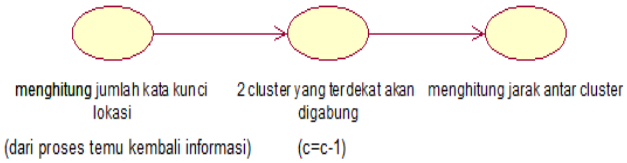
Gambar 3.4 adalah *use-case diagram* untuk proses temu kembali informasi di mana pengguna yang akan melakukan pencarian dokumen harus memasukkan kata kunci lokasi terlebih dahulu kemudian sistem akan melakukan proses temu kembali informasi terhadap dokumen sumber yang telah dicari.



Gambar 3.4 Use Case Diagram Proses Temu Kembali Informasi

3.2.2.3. Use Case Diagram Clustering dengan CLHM (*Centroid Linkage Hierarchical Method*)

Gambar 3.5 menunjukkan proses *clustering* dengan menggunakan metode CLHM (*Centroid Linkage Hierarchical Method*). Kata kunci lokasi yang dimasukkan oleh pengguna akan dicari jumlahnya oleh sistem pada dokumen kemudian jumlah ini yang akan menentukan proses *clustering* berikut.

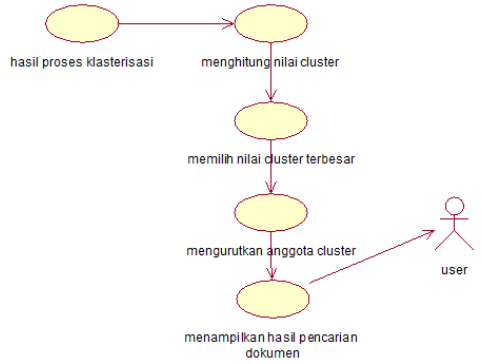


Gambar 3.5 Use Case Diagram Proses *Clustering* Dengan CLHM

3.2.2.4. Use Case Diagram Hasil Pencarian Dokumen

Gambar 3.6 menunjukkan hasil akhir dari proses clustering ini. Di mana pada akhir dari proses clustering ini akan ditampilkan hasil kumpulan dokumen yang tepat sesuai dengan kata kunci lokasi yang telah dimasukkan oleh pengguna. Dan setelah dokumen yang dicari muncul, maka

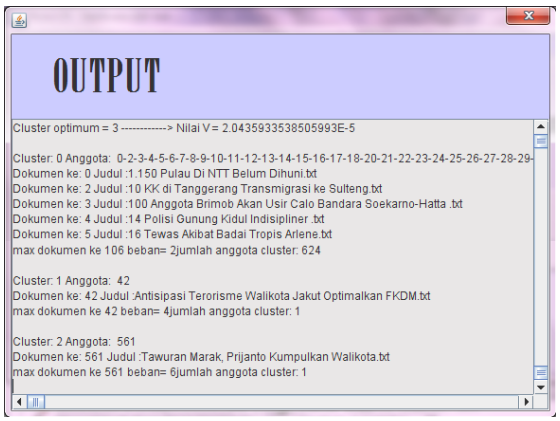
pengguna juga akan dapat langsung mengakses url asli darimana dokumen tersebut berasal.



Gambar 3.6 Use Case Diagram Proses Pencarian Dokumen Sesuai Kata Kunci Lokasi

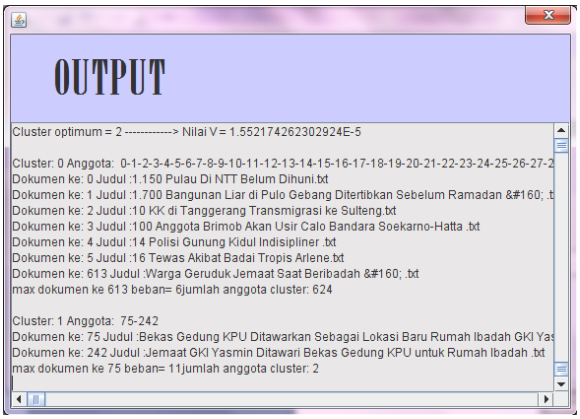
4. UJI COBA DAN ANALISA

Aplikasi Temu Kembali Informasi Berdasarkan Lokasi Pada Dokumen Yang Dikelompokkan Menggunakan Metode *Centroid Linkage Hierarchical* ini diujicobakan untuk 2 kata kunci dengan jumlah data 600 dokumen. Dan hasilnya adalah sebagai berikut:



Gambar 4.1 Hasil dokumen kata kunci jakarta

Gambar 4.2 Hasil dokumen kata kunci jakarta tangerang



Gambar 4.3 Hasil dokumen kata kunci bandung

Analisa:

- Ketika dimasukkan kata kunci lokasi bandung, dilakukan pengecekan apakah pada direktori hasil dari proses translasi (salah satu tahap modul IR) terdapat dokumen yang memiliki kata lokasi bandung maupun jakarta. Jika ada maka akan ditampilkan semua dokumen yang memiliki kata kunci lokasi bandung pada hasil translasinya.
- Cluster yang memiliki nilai variance (vb/vw) paling kecil merupakan cluster optimal yang terbentuk.
- Semakin besar bobot kata kunci yang dimiliki oleh suatu dokumen, semakin tinggi kedekatakannya dengan kata kunci lokasi yang dimasukkan oleh pengguna.

5. PENUTUP

5.1 Kesimpulan

Dari hasil percobaan serta analisis diatas, dapat diambil beberapa kesimpulan :

1. Hasil akhir dari temu kembali informasi teks berdasarkan kata kunci lokasi yang dikelompokkan dengan menggunakan metode *centroid linkage hierarchical* ini dapat digunakan untuk melakukan pencarian terhadap kata kunci lokasi yang dimasukkan oleh pengguna.
2. Hasil akhir dari temu kembali informasi teks berdasarkan kata kunci lokasi yang dikelompokkan dengan menggunakan metode *centroid linkage hierarchical* ini dapat melakukan pengklasteran paling banyak 2 kata kunci lokasi saja.
3. Semakin tinggi bobot suatu kata kunci lokasi maka semakin memiliki kedekatan atau kemiripan dengan inputan.

5.2 Saran

1. Pada proses modul IR dapat dicoba dengan menggunakan bahasa lain seperti bahasa Inggris ataupun Arab yang memiliki struktur *morphological* yang lebih kompleks daripada bahasa Inggris.
2. Pada proses pencarian dan penyimpanan dokumen secara *online* masih terbatas pada *file* html yang diunduh dari Rss berita *online* di internet, diharapkan pengembangan berikutnya tidak hanya berupa *file* html saja, namun bisa bermacam *file* seperti: pdf, ppt, doc, dll. Sehingga *clustering*nya tidak hanya terbatas pada dokumennya saja namun diharapkan bisa juga *clustering* untuk *type* data dokumen sekaligus.
3. Filter untuk proses modul IR dapat lebih diperbanyak lagi untuk menghasilkan kata kunci lokasi yang memiliki akurasi lebih tinggi.

.DAFTAR PUSTAKA

[1] Barakbah, A.R., Arai, K., *A New Algorithm For Optimization Of K-Means Clustering With Determining Maximum Distance Between Centroids*, In. IES 2006, Politeknik Elektronika Negeri Surabaya, ITS.

[2] Cahyono Dwi, Fadlil Junaidillah, Sumpeno Suryo, Hariadi Mochamad, 2008, *Temu Kembali Informasi Untuk Pembangkitan Basis Pengetahuan dari Teks Bebas yang Digunakan Oleh Agen Percakapan Bahasa Alami*, SESINDO2008.

[3] Eldira, Hervilorra, *Web Mining Untuk Pencarian Dokumen Bahasa Inggris Menggunakan Hill Climbing Automatic Cluster* 2010, Politeknik Elektronika Negeri Surabaya,ITS.

[4] Badrullami, Moh, *Rancang Bangun Aplikasi Server Crawling Berita Online Sebagai Penyedia Berita Up To Date Pada Handphone Yang Mendukung WAP*, 2010, Politeknik Elektronika Negeri Surabaya,ITS.