Visualisasi Similaritas Topik Penelitian dengan Pendekatan Kartografi Menggunakan Self-Organizing Maps (SOM)

Budi Pangestu, Diana Purwitasari dan Chastine Fatichah
Depatemen Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember (ITS)
e-mail: diana @if.its.ac.id

***Abstrak*—Penelitian merupakan salah satu hal yang penting dalam pengembangan bidang keilmuan sehingga dinilai perlu diciptakan sebuah visualisasi Peta Keterkaitan Antar Topik Riset Penelitian, agar mampu memberikan ide dan gambaran bagi calon peneliti dari Indonesia tentang potensi Topik Penelitian yang dapat dikembangkan. Pada penelitian kali ini, akan digunakan Data Penelitian Studi dari Resits.its.ac.id sebagai data input. Pemrosesan Data Mining pada data teks seringkali memiliki kendala dalam kata-kata yang terdapat pada corpus terlalu kotor atau biasa disebut stopwords, dan besarnya dimensi fitur yang didapat dari data teks sangat besar. Maka dari itu, perlu dilakukan preprocessing pada data teks yang digunakan meliputi Stopwords Removal, dan Tokenizing. Setelah melalui preprocessing, dilakukan Ekstraksi Fitur menggunakan Term Frequency – Inverse Document Frequency (TF-IDF). Reduksi fitur menggunakan Principal Component Analysis (PCA) dikenakan guna mereduksi fitur dari Data Input yang dinilai terlalu banyak. Setelah itu, Data Input dilatih dan dipetakan ke dalam 2 dimensi menggunakan metode Unsupervised Learning Self-organizing Maps (SOM). Terakhir, Teknik Clustering Kombinasi K-means dan Hierarchical Clustering dikenakan pada Jaringan Peta SOM guna mengelompokkan neuron-neuron yang terbentuk. Hasil akhir dari penelitian ini ilaha Peta Visualisasi Similaritas Topik Penelitian. Berdasarkan hasil uji coba, dapat disimpulkan bahwa ekstraksi fitur dan Teknik cluster yang digunakan sudah tepat divalidasi dengan Silhouette Score sebesar 0.5215, dan Cophenet Correlation Coefficient sebsar 0.977. Uji coba diatas menunjukkan bahwa K-means Clustering yang digunakan menghasilkan Cluster yang Cohesive dan Separable ditandai dengan hasil Silhouette Score dan Cophenet Correlation Coefficient yang besar.**

***Kata Kunci*—Hierarchical Clustering; K-means Clustering; Self-organizing Maps; Term Frequency – Inverse Document Frequency (TF-IDF); Text Mining; Visualisasi.**

I. PENDAHULUAN

PEMRINTAH memberikan perhatian penuh terutama kepada para kalangan akademisi untuk melakukan penelitian, seperti dukungan dana serta lomba-lomba keilmiahan. Kegiatan ekstrakurikuler keilmiahan juga dikembangkan mulai pendidikan tingkat menengah hingga perguruan tinggi. Sebagai salah satu perguruan tinggi di Indonesia, Institut Teknologi Sepuluh Nopember (ITS) Surabaya dengan para peneliti yang ada di dalamnya aktif memberikan kontribusi terhadap dunia penelitian Indonesia melalui publikasi jurnal dan seminar penelitian secara rutin setiap tahunnya. [1] Para peneliti dalam kapasitasnya sebagai penyedia iptek harus turut serta berperan dalam inovasi nasional. Kegiatan penelitian/riset selama ini sering terjadi antara satu dengan lain tidak ada keterkaitan. Perlu diusahakan agar kegiatan penelitian dapat dilakukan secara holistik, lebih fokus, lebih kontekstual dan ada kerjasama antar-peneliti dalam penentuan topik penelitian. [2] Selain itu, menurut survey yang dilakukan Menteri Riset, Teknologi dan Pendidikan Tinggi (Menristekdikti) jumlah publikasi yang ditelurkan peniliti Indonesia masih sangat sedikit. Pada survey yang dilakukan per Maret 2016, tercatat hanya 4.500 hingga 5.500 karya yang berhasil dipublikasikan. Jumlah tersebut tergolong kecil jika dibandingkan dengan jumlah penduduk Indonesia yang mencapai 250 juta jiwa.

Saat ini ITS sudah memiliki Sistem Repositori Peneliti. Sistem Repositori Peneliti merupakan sistem informasi yang secara khusus memberikan informasi kepada masyarakat seputar dunia penelitian yang ada di ITS. Beberapa fitur yang terdapat dalam sistem tersebut yaitu pengguna dapat melakukan pencarian peneliti dengan kriteria tertentu, melihat daftar publikasi jurnal penelitian terakhir, serta fitur lainya. Pada sistem informasi tersebut pengguna dapat melakukan pencarian peneliti berdasarkan pengelompokan area peneliti (fakultas). Pada sistem informasi tersebut juga sudah memiliki visualisasi data kerjasama peneliti dalam bentuk graph yang menarik dan mudah dipahami. Namun, Sistem Repositori Peneliti ini belum memiliki visualisasi peta yang mampu menggambarkan keterkaitan topik antar disiplin ilmu. Oleh karena itu dalam studi ini akan dibuat sebuah modul yang akan menjadi bagian dari fitur Sistem Informasi Repositori Peneliti. Modul yang akan dibuat ini akan berfokus pada visualisasi peta keterkaitan topik antar disiplin ilmu. Dengan adanya visualisasi tersebut, keterkaitan antar disiplin ilmu dapat ditampilkan secara informatif dan menarik.

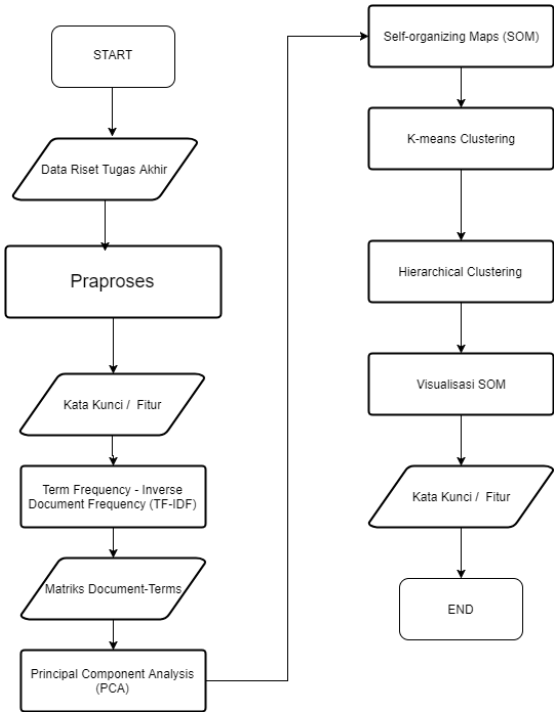
Data teks seringkali memiliki jumlah fitur yang sangat banyak, sedangkan pada Studi ini Data Dokumen Penelitian ingin divisualisasikan ke dalam peta 2 dimensi. *Self-organizing Maps* (SOM) adalah salah satu tipe dari jaringan saraf tiruan yang dilatih menggunakan *unsupervised learning* untuk mendapatkan representasi data dengan dimensi yang sedikit [3]

. SOM merupakan alat bantu yang efisien untuk merepresentasikan data dengan dimensi yang tinggi direduksi menjadi 2 dimensi saja [4]. SOM merupakan algoritma jaringan saraf tiruan yang umum digunakan untuk permasalahan ini.

Dalam studi ini akan digunakan data dokumen penelitian yang tersedia pada Sistem Repositori Peneliti ITS. Kemudian ditentukan *keyword* dari tiap dokumen tersebut yang akan dimasukkan ke dalam *vector-space model*. Kemudian, dari pemodelan tersebut dapat diketahui kemiripan antar dokumen dengan mengamati frekuensi munculnya kata-kata yang sama antar dokumen. Data *vector-space model* dokumen ini digunakan sebagai input untuk membangun *Self-organizing Maps.* Terakhir, dilakukan *K-means Clustering* dan *Hierarchical Clustering* terhadap map yang telah dibentuk untuk menyederhanakan visualisasi peta tersebut.

Harapan yang ingin dicapai dalam studi ini, para peneliti, mahasiswa ataupun masyarakat umum dapat mengetahui keterkaitan topik antar disiplin ilmu sehingga dapat digunakan sebagai acuan dalam pembuatan penelitian di masa yang akan datang.

II. METODOLOGI



Gambar 1. Gambaran umum alur dan metode yang digunakan dalam penelitian ini.

A. *Praproses*

Praproses adalah suatu proses/langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas (input yang baik untuk proses data mining). Berikut adalah langkah-langkah dalam praproses:

1) Pengambilan Data Mentah Resits.its.ac.id

Proses pertama yang dilakukan adalah mengelola data mentah yang didapat dari *repository* riset Studi milik Institut Teknologi Sepuluh Nopember menjadi data yang dapat digunakan pada sistem Visualisasi Topik Penelitian menggunakan *Self-organizing Maps* (SOM) dan Kombinasi *K-means Clustering* dan *Hierarchical Clustering*. Data *repository* ini sudah tersimpan dalam *database* dan berformat SQL. Langkah-langkah untuk mengambil data tersebut ke dalam Sistem adalah sebagai berikut:

A) Gunakan *library SQL Connector for Python* untuk menghubungkan *Python* kepada *database MySQL*.
B) Lakukan *query* untuk mengambil kolom Judul dan Abstrak dari data yang tersimpan.
C) Inisialisasi sebuah *dictionary* untuk meyimpan indeks dokumen, judul, dan abstrak dari dokumen tersebut..

2) Tokenisasi Data

Proses berikutnya adalah Judul dan Abstrak dari data yang didapat pada langkah 1 ditokenisasi guna mempermudah proses berikutnya. Tokenisasi adalah proses untuk memisahkan sebuah kalimat menjadi bagian unit yang lebih kecil atau kata-kata. Hasil dari Langkah berikut adalah kumpulan kata-kata dari kalimat pada Judul dan Abstrak.

$$Token_{judul} = \begin{bmatrix} "metode" \\ "segmentasi" \\ "jaringan" \\ "konstruksi" \end{bmatrix} \quad (1)$$

$$Token_{abstrak} = \begin{bmatrix} "metode" \\ "citra" \\ "analisis" \\ "manet" \end{bmatrix} \quad (2)$$

3) Stopwords Removal

Proses berikutnya yaitu melakukan *stopwords removal* untuk membersihkan data dari kata-kata yang menjadi noise pada proses *training*. Daftar kata-kata yang termasuk *stopwords* merupakan kata-kata penghubung yang diambil dari Kamus Besar Bahasa Indonesia dan penambahan kata-kata manual yang sering diamati muncul dan tidak bermakna pada *corpus* yang digunakan. Kata-kata yang telah ditokenisasi dicek apakah termasuk dalam daftar *stopwords*, jika kata tersebut meruapakan *stopwords*, maka kata tersebut akan dihilangkan dan tidak dipakai. Hasil dari langkah ini adalah sebagai berikut.

$$Token_{judul} = \begin{bmatrix} "metode" \\ "segmentasi" \\ "jaringan" \\ "konstruksi" \end{bmatrix} \quad (1)$$

$$Token_{abstrak} = \begin{bmatrix} "metode" \\ "citra" \\ "analisis" \\ "manet" \end{bmatrix} \quad (2)$$

B. *Ekstraksi Fitur menggunakan Term Frequency – Inverse Document Frequency (TF-IDF)*

Term Frequency – Inverse Document Frequency (TF-IDF) merupakan salah satu metode pembobotan fitur yang umum digunakan dalam mengolah data teks [5]. Pada dasarnya, TF-IDF menghitung frekuensi suatu kata muncul dalam satu

dokumen dibandingkan dengan proporsi *inverse* dari frekuensi kata tersebut muncul pada *corpus* dokumen keseluruhan. Dengan pembobotan ini, kita dapat menghitung seberapa relevan kata tersebut dengan dokumen terkait. Kata-kata yang sering muncul pada sedikit dokumen akan memiliki nilai yang lebih tinggi dibandingkan dengan kata-kata yang sering muncul pada banyak dokumen.

Jika terdapat *corpus* dokumen D, sebuah kata w, dan sebuah dokumen d ϵ D, rumus TF-IDF adalah sebagai berikut :

$$wd = fw,d * log\ (|D|/fw,D) \quad (3)$$

dimana, wd adalah bobot kata w pada dokumen d, fw,d adalah frekuensi kemunculan kata w pada dokumen d, $|D|$ adalah jumlah total dokumen pada *corpus* D, dan fw,D adalah frekuensi kemunculan kata w pada *corpus* D. Langkah-langkah dalam proses perhitungan TF-IDF adalah sebagai berikut:

1) Perhitungan TF

Untuk setiap term pada Token Abstrak untuk tiap dokumen dicek pada *list* kata kunci. Tiap *occurrence term* tersebut pada dokumen ke i, nilai TF *term* tersebut pada dokumen ke i ditambahkan.

2) Perhitungan Document Frequency

Dilakukan perhitungan *Document Frequency* dengan cara mengecek kemunculan untuk tiap *term* yang ada pada *list* kata kunci muncul pada dokumen dalam *corpus.* Tiap *term* tersebut ditemukan pada dokumen ke i, nilai *Document Frequency* ditambahkan.

3) Perhitungan Inverse Document Frequency

Mengubah hasil *Document Frequency* dari langkah 2 menjadi *Inverse Document Frequency* dengan rumus $log\ (|D|/fw,D)$ dimana *fw,D* adalah *Document Frequency* dari term *w* dan $|D|$ adalah jumlah dokumen pada *corpus D*. Berikut adalah contoh hasil dari proses TF-IDF.

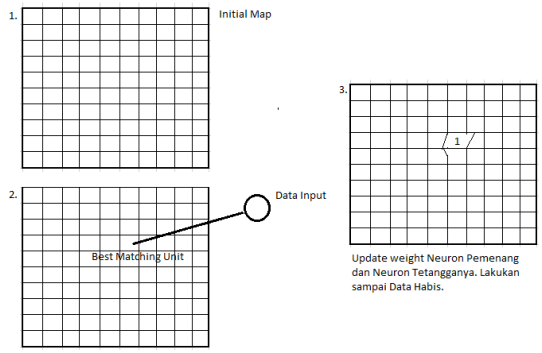
$$TF-IDF_{dokumen1} = \begin{bmatrix} analisis, 2.47712 \\ jalan, 2.17609 \\ informasi, 0.17609 \end{bmatrix}$$

$$TF-IDF_{dokumen2} = \begin{bmatrix} teknologi, 2.47712 \\ jalan, 0 \\ informasi, 3.17609 \end{bmatrix}$$

C. Self-organizing Maps (SOM)

Self-organizing Maps (SOM) adalah metode *unsupervised learning* yang digunakan pada Studi ini. Tujuan utama penggunaan SOM adalah kemampuannya untuk merepresentasikan data dengan dimensi fitur yang besar menjadi 2 dimensi saja, sehingga lebih mudah divisualisasikan.

Secara garis besar, tahapan *training SOM* dibagi menjadi 2 tahap yaitu:



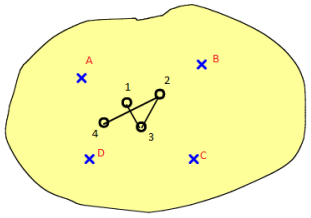
Gambar 2. Visualisasi metode *training* jaringan SOM

1. Rough Training: Fase training awal dimana weight dari tiap neuron dan juga nilai Mean Square Error (MSE) masih belum konvergen.
2. Fine Tuning: Fase training akhir dimana weight dari tiap neuron dan juga nilai Mean Square Error (MSE) sudah konvergen. Guna dari fase ini adalah menstabilkan bentuk jaringan yang sudah terbentuk.Modifikasi Penghitungan TF

Setiap proses *training SOM* dibagi kedalam 3 tahap [6] yaitu:

1. Inisialisasi
2. Sampling dan *Matching*
3. *Updating*

1. Diilustrasikan ada 4 dokumen input dengan 20 fitur kata kunci. Inisialisasi Jaringan Peta *Neuron* dengan *size* 2x2. Maka terbentuk 4 *Neuron* dengan masing-masing *Neuron* memiliki *weight* berdimensi 20 juga. Inisialisasi *Weight Neuron* berdasarkan *Eigenvectors* dengan *Eigenvalue* tertinggi dari PCA. Radius Ketetanggan Awal dinisialisasi dengan nilai 2.



Gambar 3. Visualisasi inisialisasi jaringan SOM

2. Lakukan perhitungan *Neuron Pemenang* terhadap Data Input Dokumen C

$$PCA_{dokumenC} = \begin{bmatrix} PC\ 1, 3.5672 \\ PC\ 2, 4.5481 \end{bmatrix}$$

$$d(P,Q) = \sqrt{\sum_{j=1}^{n} (FiturDataInput(j) - FiturNeuron(j))^2}$$

$$d(P,Q) = \sqrt{(3.5672 - 1.7498)^2 + (4.5481 - 2.3921)^2}$$

$$d(P,Q) = 2.963$$

Didapatkan *Neuron Pemenang* untuk Data Input C adalah *Neuron C* dengan jarak 2.963. Maka *Weight* dari *Neuron Pemenang* dan Tetangganya *diupdate*. Karena Radius ketetanggan saat ini adalah 2, maka hanya *neuron* 1 dan 2 yang dianggap tetangga dari *neuron* 3.

III. UJI COBA

A. Penghitungan Map Size Optimal pada konfigurasi SOM

Self-organizing Maps(SOM) adalah metode jaringan *Kohonen Neural Network* yang bekerja dengan cara memetakan data *input* dengan dimensi yang besar ke dalam 2 dimensi (*Many to few mapping)*. Salah satu cara yang umum digunakan untuk metode *many to few mapping* adalah perhitungan *Quantization Error*. *Quantization Error* adalah besaran yang merupakan perhitungan dari selisih dari data *input* yang asli terhadap data dengan *value* yang sudah disesuaikan pada dimensi peta SOM yang digunakan atau biasa disebut dengan *round-off error.* Hasil uji coba untuk tiap percobaan nilai *Map Size* dapat dilihat pada tabel dibawah.

Tabel 1.
Nilai *Quantization Error* pada tiap percobaan Map Size

Map Size	Quantization Error
25 x 25	49.641
40 x 40	19.752
50 x 50	8.641
75 x 75	11.264

(5)

Kesimpulan dari hasil uji coba pada Tabel 1 adalah Peta berukuran 50x50 paling optimal dikenakan terhadap data *input* sebanyak 13.324 Dokumen dengan *Quantization Error* 8.641.

B. Perhitungan Silhouette Score pada cluster yang terbentuk dari K-means

Skenario Uji Coba ketiga ialah perhitungan performa menggunakan Silhouette Scoring pada percobaan jumlah cluster pada metode K-means Clustering. Silhouette Scoring memiliki rentang nilai dari -1 sampai 1, dimana nilai negatif umumnya menandakan bahwa prediksi data untuk cluster tersebut banyak yang tidak tepat. Sedangkan nilai 0 menandakan bahwa cluster yang terbentuk masih ambigu, dimana tiap cluster yang terbentuk masih memiliki tingkap kesamaan yang tinggi antar cluster dan nilai positif jika cluster yang dibentuk sudah sesuai dengan data input dan memiliki jarak yang rendah antar tiap data pada cluster yang sama (Cohesivity) serta cluster-cluster yang terbentuk memiliki jarak yang jauh antar sesama cluster (Separability). Percobaan dilakukan terhadap 4 jumlah *cluster:*

1. 10 *Cluster* dengan *Silhouette Score* 0.4797
2. 12 *Cluster* dengan *Silhouette Score* 0.5215
3. 15 *Cluster* dengan *Silhouette Score* 0.419
4. 20 *Cluster* dengan *Silhouette Score* 0.00525

Berdasarkan uji coba diatas dapat disimpulkan bahwa nilai *Silhouette Score* tertinggi didapatkan dengan jumlah 12 *Cluster*.

IV. KESIMPULAN DAN SARAN

Kesimpulan yang dapat diambil didapatkan berdasarkan hasil uji coba Visualisasi Similaritas Topik Penelitian dengan Pendekatan Kartografi menggunakan *Self-organizing Maps* (SOM) adalah sebagai berikut:

1. Metode TF-IDF dan *Self-organizing Maps* dapat digunakan sebagai metode ekstraksi fitur dan *unsupervised learning* yang baik untuk data teks ditunjukkan dengan hasil nilai *Cophenet Correlation Coefficient* yang tinggi yaitu 0.977.
2. Berdasarkan hasil uji coba, metode *K-means Clustering* pada jaringan SOM yang sudah terlatih menghasilkan nilai rata-rata terbesar untuk penggunaan 12 jumlah *cluster*. Nilai *Silhouette Score* yang dihasilkan untuk jumlah *cluster* 12 adalah 0.5215
3. Implementasi *Term Frequency* (TF) dianggap lebih cocok untuk menentukan label dari tiap *cluster* dibandingkan *Term Frequency – Inverse Document Frequency* (TF-IDF) dikarenakan label yang dihasilkan lebih merepresentasikan tiap karakteristik dari *cluster*.
4. Teknik *K-means Clustering* dan *Hierarchical Clustering* dinilai mampu menyderhanakan jaringan peta SOM yang telah terbentuk sehingga lebih mudah divisualisasikan. Uji Validasi algoritma *clustering* diatas menggunakan *Silhouette Scoring* dengan nilai 0.5215 dan *Cophenet Correlation Coefficient* dengan nilai 0.977.
5. Peta Visualisasi Similaritas Topik Penelitian yang dihasilkan dinilai sudah memberikan makna dan informasi yang jelas dengan 73% Responden menyatakan demikian. Peta Visualisasi yang terbentuk juga dinilai sudah menarik untuk dilihat oleh 81% Responden.
6. Metode *Hierarchical Clustering* dinilai bermanfaat oleh 100% Responden dalam memberikan informasi tambahan pada peta yang telah terbentuk.
7. Peta Visualisasi yang terbentuk dari data penelitian rumpun yang lebih spesifik dinilai memberikan separasi antar topik lebih jelas oleh 69% dari Responden. Hal ini menunjukkan bahwa metode *Self-organizing Maps* memiliki performa yang lebih baik ketika dataset memiliki lebih sedikit variasi.

Saran dari penulis untuk penelitian ini ialah, perlu diimplementasikan *Rapid Automatic Keyword Indexing* sebagai metode pelabelan *cluster* yang lebih informatif

DAFTAR PUSTAKA

[1] D. (Direktorat J. P. Tinggi), "http://www.dikti.go.id/kolokium-di-australia-kerjasama-antar-penelitisemakin-dibutuhkan-di-indonesia/." .

[2] K. R. dan T. R. Indonesia, "http://www1.ristek.go.id/?module=News%20News&id=8705," *ristek.go.id*. [Online]. Available: http://www1.ristek.go.id/?module=News News&id=8705.

[3] A. Skupin, "A Cartographic Approach to Visualizing Conference Abstracts," 2013.

[4] J. A. Bullinaria, *Self Organizing Maps: Fundamentals*. 2004.

[5] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," 2007.

[6] dan J. O. I. Valova, G. Georgiev, N. Gueorguieva, "Initialization Issues in Self-organizing Maps," *Sci. Direct*, 2013.