Pengkategorian Otomatis Artikel Ilmiah dalam Pangkalan Data Perpustakaan Digital Menggunakan Metode Kernel Graph

Automatic Categorization of Scientific Article in Digital Library Database Using Graph Kernel Method

Budi Nugroho[1], Ekawati Marlina[2]
[1&2]Pusat Dokumentasi dan Informasi Ilmiah Lembaga Ilmu Pengetahuan Indonesia

[1]budi.nugroho@lipi.go.id, [2]ekawati.marlina@lipi.go.id

Naskah diterima: 3 Oktober 2017, direvisi: 5 Desember 2017, disetujui: 6 Desember 2017

Abstrak

Artikel ilmiah dalam pangkalan data perpustakaan digital dikelompokkan dalam kategori-kategori tertentu. Pengelompokan artikel ilmiah dalam jumlah besar yang dilakukan secara manual membutuhkan sumber daya manusia yang banyak dan waktu yang tidak singkat. Penelitian ini bertujuan untuk membantu tim pengolah bahan pustaka dalam mengelompokkan artikel ilmiah sesuai dengan kategorinya masing-masing secara otomatis. Dalam penelitian ini, pengkategorian otomatis artikel ilmiah dilakukan dengan menggunakan kernel graph yang diterapkan pada graph bipartite antara dokumen artikel ilmiah dengan kata kuncinya. Lima fungsi kernel digunakan untuk menghitung nilai matriks kernel, yaitu KE_{Gauss}, KE_{Linear}, KV_{Gauss}, KV_{Linear} dan KRW. Matriks kernel dihitung dari proyeksi satu-moda graph bipartit, lalu digunakan sebagai masukan pengklasifikasi SVM (support vector machine) dalam menentukan kategori yang tepat. Kinerja pengkategorian otomatis dihitung dari ketepatan yang merupakan perbandingan antara jumlah artikel yang dikategorikan secara tepat dengan jumlah keseluruhan artikel dalam dataset. Penerapan metode ini dalam pangkalan data ISJD (Indonesian Scientific Journal Database) menghasilkan rata-rata ketepatan yang signifikan yaitu 87,43% untuk fungsi kernel KV_{Gauss}. Sedangkan kernel lainnya memberikan hasil berturut-turut 86,14% (KE_{Linear}), 85,86% (KE_{Gauss}), 42,23% (KV_{Linear} dan 25,15% (KRW). Hasil ini menunjukkan bahwa penggunaan metode kernel graf efektif untuk mengelompokkan artikel ilmiah ke dalam kategori yang ditentukan dalam pangkalan data perpustakaan digital.

Kata kunci: *automated classification, graph kernel, bipartite graph, support vector machine, digital library database*

Abstract

Academic paper in a digital library database are classified into particular categories by doing such classification task. A manual classification task is time consuming and labor intensive. This study aims to assist information specialists in classifying academic paper automatically. We employed graph kernel method to bipartite graph of academic paper document with their keywords. We used five kernel functions, namely KE_{Gauss}, KE_{Linear}, KV_{Gauss}, KV_{Linear} and KRW to calculate kernel matrix. The kernel matrix was calculated from one mode projection of the bipartite graph and used as input for SVM (support vector machine)

classifier to classify academic paper into appropriate categories. The performance of SVM classifier was calculated by accuracy which is the comparison between the number of correctly classified instances per category and the number of instances. We applied the method to ISJD (Indonesian Scientific Journals Database) dataset. We obtained an average of accuracy 87,86% from KV_{Gauss} kernel function. While other kernels yielded results 86,14% (KE_{Linear}), 85,86% (KE_{Gauss}), 42,23% (KV_{Linear} and 25,15% (KRW) respectively. The results showed that using graph kernel method is effective for classifying academic paper into such categories defined in digital library database.

Keywords: *automated classification, graph kernel, bipartite graph, support vector machine, digital library database*

PENDAHULUAN

Artikel ilmiah merupakan sumber informasi ilmiah yang sekaligus dapat memperlihatkan perkembangan ilmu pengetahuan dan teknologi. Untuk sampai dapat diakses dan dinikmati oleh pengguna informasi, artikel ilmiah mengalami proses pengolahan literatur terlebih dahulu sehingga bisa ditampilkan dalam pangkalan data perpustakaan digital. Proses pengklasifikasian artikel ilmiah dilakukan oleh petugas pengolah bahan pustaka (pustakawan) secara manual. Pustakawan membaca bahan pustaka yang diolah, dengan alat bantu berupa *tesaurus*/daftar kata kunci berbagai bidang dan LC subject headings, mereka menentukan indeks kata kunci, kelas DDC dan kategori bidang ilmu. Seiring dengan meningkatnya jumlah artikel ilmiah yang akan diolah dan keterbatasan sumber daya yang mengerjakan, proses pengklasifikasian ini memakan waktu yang tidak singkat. Untuk mengatasi kendala ini, pengategorian otomatis artikel ilmiah perlu dilakukan.

Penelitian mengenai pengategorian otomatis teks atau dokumen berdasarkan metode *kernel graph* telah dilakukan dengan memodelkan artikel jurnal ilmiah sebagai *graph bipartite* dengan kata kunci yang dimilikinya (Srivastava et al., 2013). menggunakan proyeksi satu-moda dan algoritma optimasi modularitas untuk mengelompokkan (*clustering*) dokumen dalam suatu pangkalan data. Masih dalam topik *clustering*, (Dhillon, 2001) memanfaatkan algoritma *co-clustering* spektral untuk memodelkan koleksi dokumen sebagai *graph bipartite* antara dokumen dengan kata dan menyelesaikan persoalan pengelompokkan simultan. Penelitian lain yang bertujuan untuk pengategorian otomatis dilakukan oleh Banerjee et al. (2012) untuk mengkategorikan individu atau organisasi berdasarkan kepercayaan, keseharian dan ekspresi pendapat yang dimodelkan dalam *graph bipartite*. Penelitian tersebut memiliki perhatian yang sama dengan penelitian yang dilakukan oleh (Andrej dan Doreian, 2009) yaitu fokus pada pengategorian menggunakan *graph bipartite*. (de Paulo Faleiros et al., 2016) mengklasifikasikan dokumen teks menggunakan propagasi pada *graph bipartite*.

Penggunaan *graph bipartite* untuk pengategorian artikel ilmiah dalam pangkalan data perpustakaan digital belum banyak dilakukan, kecuali (Radev, 2009) yang membandingkan dua algoritma klasifikasi (partisi spektral dan pembaharuan *tripartite*) untuk mengklasifikasi angka dalam kumpulan teks. Penelitian lainnya merupakan topik *clustering*, ialah Yoo et al. (2006), Zha et al. (2001), Grace dan Desikan (2016), dan Zha and Ji (2002).

Dalam penelitian ini, digunakan *kernel graph* yang diterapkan pada *graph bipartite*, yaitu antara dokumen artikel ilmiah dengan kata kuncinya untuk mengkategorikan artikel ilmiah secara otomatis. Gagasan

menggunakan *kernel graph* berdasarkan penelitian (Sugiyama dan Borgwardt, 2015) mengenai macam-macam kernel graph dan penelitian (Srivastava et al., 2013) tentang teknik satu-moda untuk proyeksi *graph bipartite*. Penelitian ini memberikan kontribusi dalam topik pengategorian otomatis artikel ilmiah sebagai berikut.

1. Menjelaskan macam-macam kernel graph yang bisa digunakan dalam pengkategorian otomatis artikel ilmiah yang dimodelkan ke dalam bentuk graph bipartite.
2. Menguraikan teknik matriks kernel sebagai masukan untuk pengklasifikasi SVM dalam pengkategorian otomatis artikel ilmiah.
3. Menguraikan pemanfaatan teknik machine learning untuk membantu mempercepat proses pengklasifikasian artikel ilmiah menggunakan metode kernel yang dipadukan dengan teori graph, khususnya graph bipartite.

Pengolahan Artikel Ilmiah

Artikel ilmiah dalam jurnal ilmiah Indonesia merupakan sumber informasi ilmiah yang sekaligus dapat memperlihatkan perkembangan ilmu pengetahuan di Indonesia. Pusat Dokumentasi dan Informasi Ilmiah Lembaga Ilmu Pengetahuan Indonesia (PDII-LIPI) merupakan institusi yang mengumpulkan berbagai jenis karya ilmiah, termasuk di dalamnya jurnal ilmiah. Jurnal ilmiah ini menjadi sumber informasi yang penting bagi akademisi ataupun peneliti yang ingin melakukan penelitian ataupun pengkajian. Oleh karena itu, PDII-LIPI membuka akses ke koleksi jurnal ilmiahnya untuk masyarakat umum melalui situs ISJD.

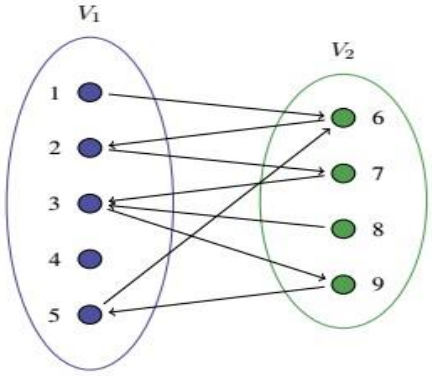
Melalui situs ini, pemustaka dapat melihat data bibliografis dan artikel lengkap dari koleksi artikel ilmiah PDII-LIPI. Namun, untuk sampai dapat diakses dan dinikmati oleh pemustaka, setiap artikel jurnal ilmiah yang diterima oleh PDII-LIPI, harus mengalami proses pengolahan literatur terlebih dahulu. Secara umum, pengolahan literatur dapat diartikan sebagai suatu alur kegiatan kerja yang berhubungan dengan mengolah koleksi bahan perpustakaan, mulai tiba di perpustakaan sampai siap untuk digunakan atau dipinjam oleh pemustaka. Artikel ilmiah dalam jurnal diolah dengan menganalisis subyek masing-masing dengan menggunakan *tesaurus* dan menentukan kategori masing-masing artikel jurnal tersebut dengan menggunakan *dewey decimal classification* (DDC) dan kategori bidang ilmu yang dimiliki oleh pangkalan data ISJD melalui aktivitas pengolahan literatur yang disebut proses pengklasifikasian.

Klasifikasi ialah kegiatan kerja mengelompokan koleksi dengan cara memberikan kode tertentu agar koleksi yang sejenis dapat terkumpul menjadi satu. Pengolahan artikel jurnal terdiri atas entri artikel, penentuan indeks kata kunci dan kelas DDC, serta validasi. Proses pengklasifikasian bahan pustaka, khususnya artikel ilmiah dilakukan oleh pustakawan secara manual. Pustakawan membaca bahan pustaka yang diolah, dengan alat bantu berupa tesaurus/daftar kata kunci berbagai bidang dan LC *subject headings*. Mereka menentukan indeks kata kunci, kelas DDC dan kategori bidang ilmu. Seiring dengan meningkatnya jumlah artikel ilmiah yang akan diolah dan keterbatasan sumber daya yang mengerjakan, proses pengklasifikasian ini memakan waktu yang tidak singkat. Hal ini mengakibatkan penumpukan bahan pustaka yang belum diolah dan akhirnya artikel ilmiah tidak dapat segera diakses oleh pengguna.

Kernel Graph

Sebuah *graph* seringkali diidentikkan dengan gambar yang diplot berdasarkan sumbu x dan sumbu y. Di bidang matematika dan ilmu komputer, *graph* merupakan konstruksi teoritis yang terdiri atas simpul-simpul (*vertices*) yang terhubung oleh sisi (*edges*). *Graph* mewakili data yang terstruktur. Simpul melambangkan satuan

diskrit informasi, sedangkan sisi melambangkan hubungan antar satuan informasi tersebut. *Graph bipartite G* yang dinyatakan sebagai $G(V_1,V_2,E)$ didefinisikan sebagai *graph* yang himpunan simpulnya bagian V_1 dan V_2, sedemikian rupa sehingga setiap sisi *E* pada *G* menghubungkan sebuah simpul di V_1 ke sebuah simpul di V_2. Kumpulan data artikel ilmiah jurnal dapat digambarkan sebagai *graph bipartite* dengan artikel ilmiah sebagai simpul-simpul V_1 dan kata-kata kunci sebagai simpul-simpul V_2, sedangkan keberadaan secara bersama-sama artikel dengan kata-kata kuncinya sebagai sisi-sisi *E*. Ilustrasi mengenai *graph bipartite* ini dapat dilihat pada Gambar 1 di bawah ini.



Gambar 1. Contoh *Graph Bipartite*

Artikel ilmiah dalam suatu pangkalan data perpustakaan digital diklasifikasikan pada kategori-kategori tertentu. Pemodelan data artikel ilmiah menjadi *graph bipartite* dapat mengklasifikasi artikel ilmiah dengan menggunakan metode berbasis kernel. Metode berbasis kernel terdiri atas fungsi kernel dan mesin kernel. Fungsi kernel menghitung ukuran kemiripan antara kumpulan data dan membuat ruang fitur. Mesin kernel adalah algoritma pembelajar (*learning algorithm*) yang menjalankan tugas pembelajar dalam ruang fitur yang dibangun oleh fungsi kernel. Selama proses ini, mesin kernel hanya memerlukan matriks kernel yang mengandung nilai kernel (nilai kemiripan) antara pasangan kumpulan data. Hal ini tidak memerlukan pendefinisian ruang fitur dan vektor fitur secara eksplisit. Dalam ruang fitur yang didefinisikan oleh fungsi kernel, macam-macam mesin kernel dapat digunakan untuk melakukan proses pembelajar, misalnya algoritma SVM. Kinerja mesin kernel sangat dipengaruhi oleh pemilihan dan desain fungsi kernel. Dalam penelitian ini, proses klasifikasi dilakukan dengan menggunakan fungsi kernel yang diadaptasi dari penelitian (Sugiyama and Borgwardt 2015).

a. Kernel Histogram Label Simpul (*Vertex Label histograms Kernel*)

Diketahui sepasang graph G_1 dan G_2. Diasumsikan bahwa rentang label $^P = \{1,2,...,s\}$ tanpa kehilangan keumumannya (*generality*). Histogram label simpul graph $G = (V,E,\phi)$ adalah vektor $f = (f_1,f_2,...,f_s)$, sehingga $f_i = |\{v \in V \mid \phi(v) = i\}|$ untuk setiap $i \in {}^P$. Ditentukan *f* dan f^0 adalah histogram label simpul graph G_1 and G_2. Kernel histogram label simpul $K_V(G_1,G_2)$ didefinisikan sebagai kernel linier antara *f* dan f^0:

$$K_V(G_1,G_2) = \langle f, f' \rangle = \sum_{i=1}^{s} f_i f_i' \quad .$$

b. Kernel Histogram Label Sisi (*Edge Label Histograms Kernel*)

Histogram label sisi adalah vektor $g = (g_1,g_2,...,g_s)$, sehingga $g_i = |\{(u,v) \in E \mid \phi(u,v) = i\}|$ untuk setiap $i \in {}^P$. Kernel histogram label sisi $KE(G_1,G_2)$ didefinisikan sebagai kernel linier antara g_1 dan g_2 untuk masing-masing histogram:

$$KE(G_1,G_2) = \langle g, g' \rangle = \sum_{i=1}^{s} g_i g_1' \quad .$$

c. Kernel *Random Walk*

Ditentukan $G = (V,E,\phi)$ adalah graph berlabel, dimana *V* adalah kesatuan simpul, *E* adalah kesatuan sisi, and ϕ adalah pemetaan $\phi : V \cup E \rightarrow {}^P$ dengan rentang ^{P} label simpul dan label sisi. Untuk sebuah sisi $(u,v) \in E$, teridentifikasi (u,v) dan (v,u) jika *G* adalah graph tak berarah (*undirected*). Derajad simpul (*degree of vertex*) $v \in V$ dinyatakan dengan $d(v)$. Hasil kali langsung (*direct*

(tensor) product) $G_x = (V_x, E_x, \phi_x)$ dari dua graph $G = (V, E, \phi)$ dan $G^0 = (V^0, E^0, \phi^0)$ didefinisikan sebagai berikut:

$$V_x = \{(v, v;) \in V \times V^0 \mid \phi(v) = \phi^0(v^0)\},$$

$$E_x = \{((u, u^0), (v, v^0)) \in V_x \times V_x \mid (u, v) \in E, (u^0, v^0) \in E^0, and \phi(u, v) = \phi^0(u^0, v^0)\},$$

dan semua label diberikan, atau $\phi_x((v, v^0)) = \phi(v) = \phi^0(v^0)$ dan $\phi_x((u, u^0), (v, v^0)) = \phi(u, v) = \phi^0(u^0, v^0)$. A_x dinyatakan sebagai matriks ketetanggaan (*adjacency matrix*) dari G_x, sedangkan δ_x dan Δ_x adalah derajad minimum dan derajad maksimum dari G_x.

Untuk mengukur kemiripan antara graph G and G^0, kernel random walk menghitung semua pasangan langkah tepat (*matching walks*) pada G and G^0. Jika diasumsikan terdapat distribusi yang seragam untuk setiap kemungkinan mulai dan berhenti di seluruh simpul-simpul G and G^0, jumlah langkah tepat diperoleh melalui matriks ketetanggaan A_x dari perkalian graph G_x. Untuk setiap $k \in N$, kernel k-step random walk antara dua graph G dan G^0 didefinisikan sebagai:

$$K_x^k(G, G') = \sum_{i,j=1}^{|V_x|} \left[\sum_{l=0}^{k} \lambda_l A_x^l \right]_{ij}$$

dengan urutan bilangan positif, bobot bernilai nyata (*real-valued weights*) $\lambda_0, \lambda_1, \lambda_2, ..., \lambda_k$ dengan asumsi $A_x^0 = I$, matriks identitas.

Pemodelan *Graph Bipartite* untuk Pengkategorian Otomatis Artikel Ilmiah

(Banerjee et al., 2012) menggunakan *graph bipartite* untuk memodelkan ekspresi pendapat antara salah satu tipe entitas (mis. individu, organisasi) dengan yang lainnya (mis. pandangan politik, keyakinan beragama), dan berdasarkan kekuatan pendapat tersebut, dibuat partisi dua tipe entitas menjadi dua klaster. Dalam penelitian ini, telah dikembangkan *tool* partisi otomatis untuk mengklasifikasi individu dan atau organisasi menjadi dua kelompok terpisah berdasarkan kepercayaan, praktik keseharian dan ekspresi pendapat. Penelitian (Banerjee et al., 2012) berfokus untuk memberikan solusi dalam permasalahan penanda partisi graph bipartite (*Signed Bipartite Graph Partition Problem (SBGPP)*. Bila sebuah sisi berlabel dan berbobot dari graph bipartite $G = (U \cup V, E)$ di mana $U = \{u_1, u_2, ..., u_n\}$ melambangkan entitas tipe I dan $V = \{v_1, v_2, ..., v_m\}$ melambangkan entitas tipe II, maka SBGPP bertujuan unutk mempartisi U ke dalam dua kesatuan terpisah U_1 dan U_2 (mirip dengan V menjadi V_1 dan V_2). Algoritma yang digunakan terdiri atas dua macam yaitu *Move-based Heuristic (MBH)* dan *Node Gain Computation*.

Penelitian mengenai pengklasteran dokumen dan pengklasteran kata telah banyak dilakukan, tetapi sebagian besar algoritma tidak dapat melakukan keduanya secara simultan. Dhillon (2001) melakukan studi pemodelan koleksi dokumen sebagai *graph bipartite* antara dokumen dengan kata, yang mana pengklasteran secara simultan dapat diselesaikan, seperti masalah partisi dalam *graph bipartite*. Dalam penelitian ini, telah digunakan algoritma *spectral co-clustering* yang memanfaatkan vektor tunggal kedua sebelah kanan dan kiri dari matriks dokumen-kata terskala untuk menghasilkan partisi yang baik.

(Radev, 2009) melakukan penelitian untuk mengklasifikasi angka dalam teks dokumen dengan memodelkannya sebagai *graph bipartite*. Klasifikasi tipe-tipe angka dilakukan dengan menggunakan algoritma partisi spektral dan pembaharuan *tripartite*. Berdasarkan hasil eksperimen, kedua algoritma ini tidak memerlukan jumlah data training yang besar untuk menghasilkan klasifikasi yang baik. Algoritma pembaharuan *tripartite* memberikan hasil klasifikasi yang lebih baik dari pada algoritma partisi spektral.

Kemiripan dokumen dapat diperoleh dengan mengambil proyeksi satu-moda dari *graph bipartite*. Srivastava et al. (2013)

mengkaji penggunaan proyeksi satu-moda dari *graph bipartite* dokumen-kata dan menerapkan algoritma optimasi modularitas untuk mengklaster dokumen. Algoritma yang diusulkan dalam penelitian ini terdiri atas dua langkah: pertama, menelusur dokumen yang mudah untuk diklaster, lalu menandai dokumen yang tersisa ke dalam klaster yang ada atau klaster baru. Algoritma ini memberikan hasil yang lebih baik daripada pendekatan clustering tradisional.

Stankova et al. (2015) membangun kerangka kerja klasifikasi tiga tahap untuk menyelesaikan permasalahan klasifikasi simpul pada dataset yang berukuran sangat besar. Pertama, penentuan bobot simpul-simpul atas. Kedua, *graph bipartite* diproyeksikan menjadi *graph unipartite* (homogen) di antara simpul-simpul bawah, di mana bobot sisi-sisi merupakan fungsi bobot simpul-simpul atas dari *graph bipartite* tersebut. Terakhir, pengklasifikasi diterapkan untuk menghasilkan *unigraph* terbobot.

METODE

Eksperimen untuk meneliti penerapan fungsi kernel dalam mengklasifikasi dokumen artikel ilmiah dilakukan dengan mengeksploitasi *graph bipartite* artikel ilmiah dengan kata kuncinya. Penggunaan kernel graph dilandasi oleh penelitian Sugiyama dan Borgwardt (2015) mengenai macam-macam kernel graph dan penelitian Srivastava et al. (2013) tentang teknik satu-moda untuk proyeksi graph bipartite. Kami menggunakan lima macam *kernel graph*, yaitu: kernel Gaussian RBF antara histogram label sisi (KE_{Gauss}), kernel linier antara histogram label sisi (KE_{Linear}), kernel Gaussian RBF antara histogram label simpul (KV_{Gauss}), kernel linier antara histogram label simpul (KV_{Linear}) dan kernel Random Walk (*KRW*). Pengklasifikasi dilatih menggunakan matriks kernel data dalam dataset training. Dalam penelitian ini, pengklasifikasi yang digunakan adalah SVM karena telah terbukti kinerjanya dalam penelitian sebelumnya (Sugiyama dan Borgwardt, 2015).

Sesuai dengan tujuan penelitian untuk mengklasifikasi dokumen artikel ilmiah, ukuran kemiripan merupakan sebuah fungsi yang berhubungan dengan nilai numerik dari pasangan *graph bipartite* dengan konsep bahwa nilai yang lebih tinggi menunjukkan kemiripan yang lebih dengan antara graph. Terdapat hubungan positif antara matriks kernel dengan matriks kemiripan berdasarkan jarak (*distance-based*). Kerangka kerja umum yang digunakan dalam penelitian ini diadaptasi dari (Martın De Diego et al., 2010), yang masing-masing berhubungan dengan tahapan *training* dan *testing* dari proses klasifikasi.

Algoritma Training

1. Ditentukan $K_1,K_2,...,K_M$ merupakan kesatuan input ternormalisasi dari matriks kemiripan yang dihitung dari poin data training $\{x_i,...,x_l\}$ digambarkan dari distribusi statistik X.
2. Buat matriks kemiripan simetris tunggal $K^* = h(K_1,K_2,...,K_M)$, di mana h adalah fungsi non linier dari matriks input dan label $\{y_i,...,y_l\}$ pada dataset training.
3. Bila perlu, transformasikan K^* menjadi matriks kernel $K_{psd}{}^*$ (*symmetric positive semi-definite matrix*).
4. Gunakan K^*_{psd} untuk melatih pengklasifikasi SVM untuk menghitung vektor bobot α yang akan digunakan untuk membuat aturan diskriminan pada saat testing.

Algoritma Testing

1. Pertimbangkan poin tak terlabel x.
2. Hitung $f_{nclass}(x)\ \sum_{i=1}^{l} \alpha_i y_i K^*_{nclass}(x, x_i)$, dimana $K^*_{nclass}(x, x_i)$ berhubungan dengan $K^*_{psd}(x, x_i)$ dengan asumsi x anggota kategori $nclass$, α adalah vektor bobot.
3. Hitung $f(x) = sign(f_{nclass}(x))$.

Dataset

Eksperimen dilakukan terhadap koleksi artikel ilmiah yang diperoleh dari pangkalan data ISJD. Kategori artikel ilmiah diambil dari kategori yang tersedia dalam dataset tersebut. *Dataset* yang digunakan dalam penelitian ini dibatasi pada rentang tahun 2012 sampai dengan 2016. Untuk keperluan eksperimen, *graph bipartite Mnet* dibangun dari 1000 data pertama dari keseluruhan *dataset*. Dokumen artikel ilmiah dibagi ke dalam dua kesatuan dengan penarikan contoh acak: (a) kesatuan training (80% dari keseluruhan *dataset*) dengan artikel ilmiah diketahui kategorinya; (b) kesatuan testing (20% dari keseluruhan *dataset*) dengan artikel ilmiah yang akan ditentukan kategorinya. Ringkasan dataset artikel ilmiah ISJD disajikan dalam Tabel 1.

HASIL DAN PEMBAHASAN

Setelah mendapatkan *dataset* dari pangkalan data ISJD, kami membuat matriks *co-occurance* kata kunci. Matriks ini selanjutnya diproses menjadi objek *igraph Mnet* (*graph bipartite*) yang kemudian dibuat proyeksi *bipartite* (*Proj*1 untuk simpul artikel, dan *Proj*2 untuk simpul kata kunci). Atribut bidang untuk pengkategorian ditambahkan pada proyeksi *Proj*1 yang akan digunakan untuk proses klasifikasi simpul dokumen artikel ilmiah. Nilai kernel KE_{Gauss}, KE_{Linear}, KV_{Gauss}, KV_{Linear}, *KRW* dengan sebelumnya membuat daftar (*list*) objek $igraph$ dari $Proj1$. Tahap selanjutnya, ditentukan kesatuan (set) *training* dan *testing* dari matriks kernel K dengan contoh acak 80% $(trainK)$, 20% $(testK)$. Pengklasifikasi SVM diterapkan pada kesatuan training $trainK$ matriks kernel dengan *k-fold cross validation* (cross = 10) untuk memperoleh model *m*. *testK* matriks kernel dari sisa 20% matriks kernel *K* ditransformasi dengan mengindeks (menggunakan SVindex) pada model *m*. Model *m* diterapkan untuk mengklasifikasi *testK* matriks kernel. Sebagai tahap terakhir, kami mengevaluasi kinerja model dengan menghitung ketepatan (*accuracy*) dan *F-measure*.

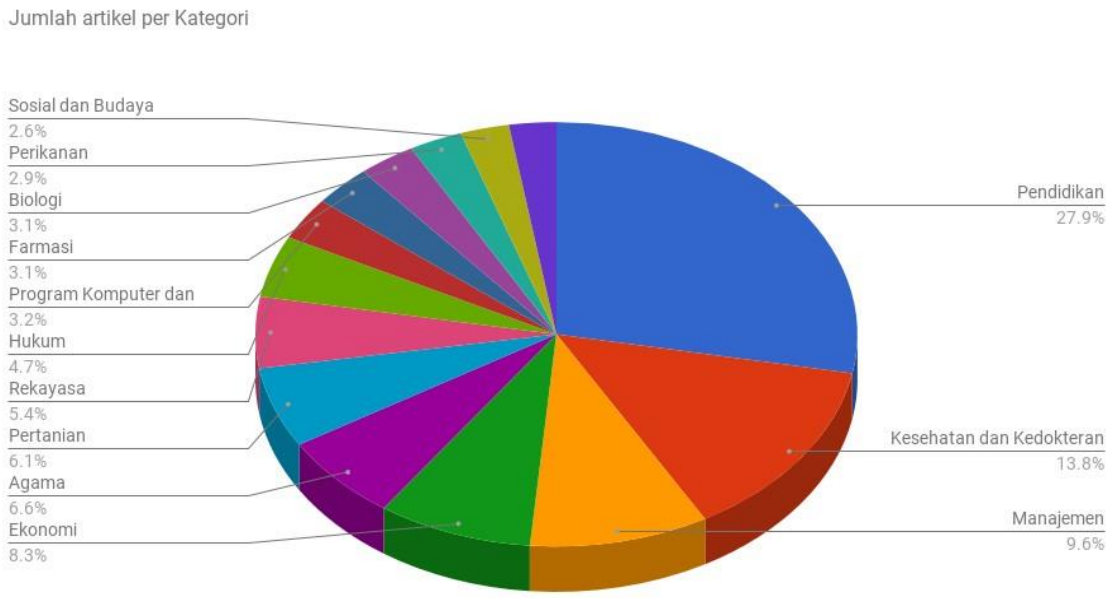
Deskripsi Dataset ISJD

Jumlah Artikel dan Kata Kunci per Kategori

Data diambil dari pangkalan ISJD pada tanggal 5 Mei 2017. Jumlah artikel ilmiah Indonesia yang terbit dari tahun 2012-2016 yaitu sebanyak 59.380 artikel yang terbagi menjadi 46 kategori. Dalam ISJD, artikel dikelompokkan menjadi kategori bidang ilmu. Penentuan jenis-jenis kategori berdasarkan dari Dewey Decimal Classification (DDC). Sebaran dari artikel berdasarkan pada kategori ditampilkan pada Gambar 2 untuk 14 kategori pertama.

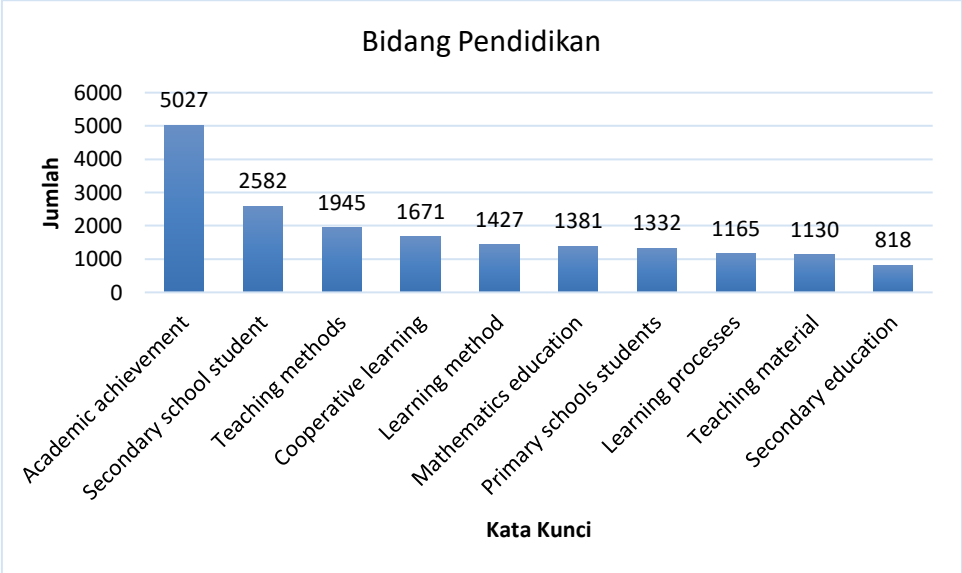
Tabel 1. Ringkasan Dataset Artikel

Dataset	$\sum V_1$	$\sum E_1$	Kategori	$\sum V_2$	$\sum E_2$
M_{net}	1000	6004	43	46270	8059

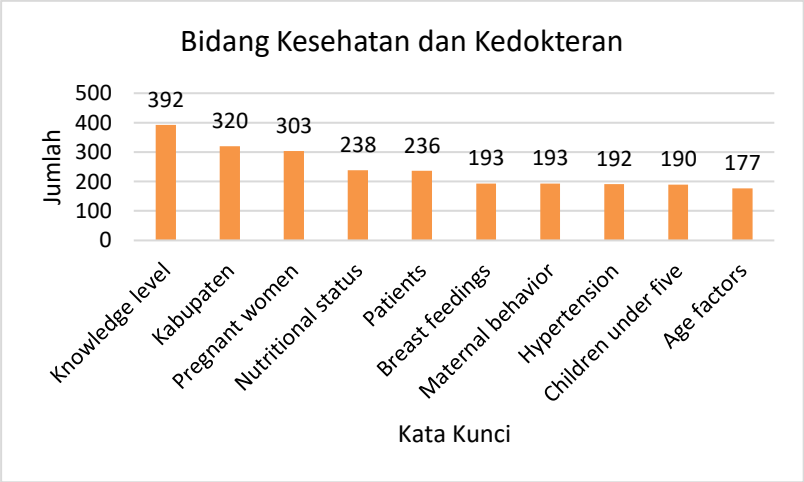


Gambar 2. Sebaran Artikel per Kategori

Dari gambar 2 di atas, dapat kita ketahui bahwa kategori terbanyak ialah pendidikan dengan jumlah artikel sebanyak 12.979 atau 21.9%. Bidang terbanyak kedua ialah bidang kesehatan dan kedokteran dengan jumlah artikel ilmiah sebanyak 6428 artikel atau 10,8%. Dari kedua kategori dengan jumlah artikel terbanyak ini, masing-masing mempunyai kata kunci yang menggambarkan karakteristik kategori tersebut, sebagaimana disajikan dalam gambar 3 dan gambar 4 berikut.



Gambar 3. Bidang Pendidikan



Gambar 4. Bidang Kesehatan dan Kedokteran

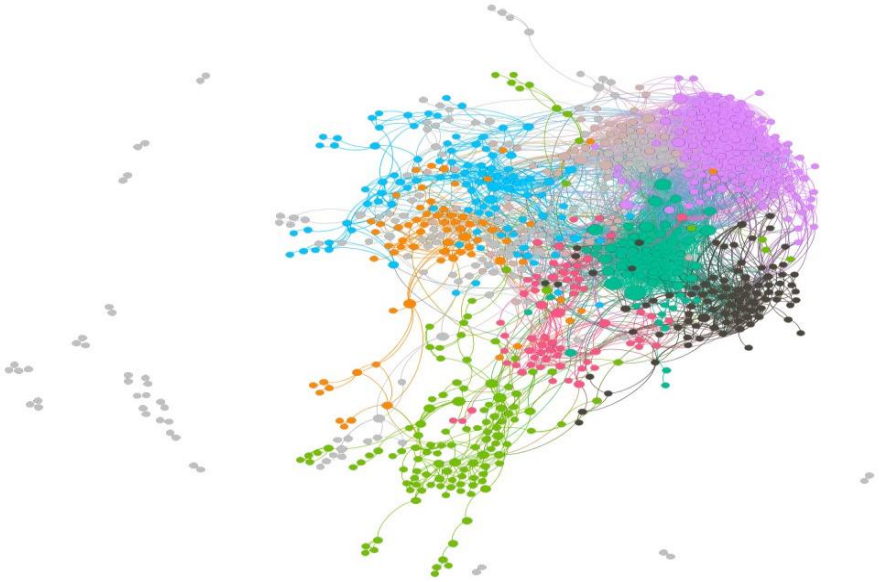
Graph Bipartite Dataset ISJD

Graph bipartite *Mnet* terdiri atas dua simpul, yaitu V_1 yang merupakan simpul artikel dan V_2 sebagai simpul kata kunci, sebagaimana diilustrasikan dalam Gambar 5. Graph *Mnet* ini mempunyai simpul artikel sebanyak 1000 dan simpul kata kunci 42270, dengan sisi-sisi masing-masing 6004 dan 8059.

Gambar 5. *Mnet* Graph Bipartite

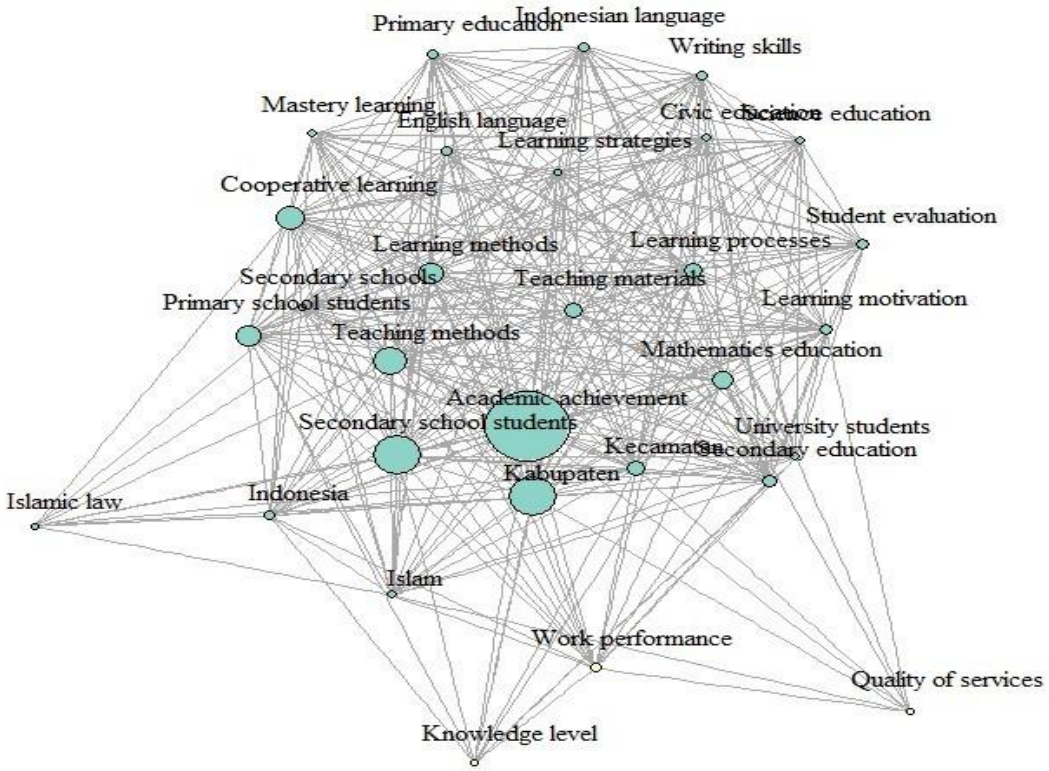
Pada tahapan proyeksi, diperoleh dua buah *graph* terpisah antara simpul V_1 dan V_2. Gambar 6 menyajikan hasil proyeksi *Proj*1 yang merupakan graph dari simpul artikel. Dari 1000 data yang dijadikan bahan eksperimen, dengan menerapkan ukuran

modularitas, data artikel ilmiah ini terklaster ke dalam 32 klaster. Hal ini sedikit berbeda dengan data awal sebagaimana tercantum dalam Tabel 1 yang mana dalam dataset ini terdapat 43 kategori (dinamakan: bidang). *Graph Proj*1 setelah ditambahkan atribut bidang, digunakan untuk proses mengategorikan artikel ilmiah.



Gambar 6. *Proj*1 Graph Bipartite

Proses proyeksi juga menghasilkan proyeksi kedua *Proj*2 yang merupakan graph simpul kata kunci. Seperti tercantum dalam Tabel 1, graph kata kunci mempunyai simpul sebanyak 46270 dan sisi 8059. Untuk memudahkan penggambaran, tampilan ilustrasi dibatasi, sebagaimana tampak dalam gambar 7 dengan jumlah simpul 30.



Gambar 7. *Proj*2 Graph Bipartite, n_{simpul} = 30

Hasil Klasifikasi Artikel Ilmiah

Proses klasifikasi dilakukan dengan menggunakan graph *Proj*1 yang telah ditambahkan label bidang masing-masing untuk simpul artikel ilmiah. Kami menggunakan kode bidang untuk membangun indeks bagi pengklasifikasi SVM dalam matriks kernel yang digunakan dengan input. Setelah proses penghitungan nilai kernel matrik dan proses klasifikasi, diperoleh hasil klasifikasi sebagaimana disajikan dalam Tabel 2 berikut.

Tabel 2. Hasil Klasifikasi Artikel Ilmiah

Kernel	Accuracy (%)	F-Measure (%)
KE_{Gauss}	85,86	47,91
KE_{Linear}	86,14	**92,75**
KV_{Gauss}	**87,43**	44,29
KV_{Linear}	42,34	19,91
KRW	25,25	25,21

Dari Tabel 2 di atas, dapat kita amati bahwa kernel KV_{Gauss} memberikan nilai ketepatan yang relatif besar dibandingkan kernel lainnya, yaitu sebesar 87,43%. Sementara itu, dua kernel lainnya yaitu KE_{Linear} dan KE_{Gauss} memberikan hasil yang masih relatif signifikan untuk proses klasifikasi, di atas Tabel 2. Hasil Klasifikasi artikel ilmiah 80%, masing-masing 86,14% dan 85,86%. Adapun kernel KV_{Linear} dan *KRW* memberikan hasil klasifikasi dengan ketepatan rendah, masing-masing 42,34% dan 25,25%.

Pengelompokan dengan fungsi kernel memperlihatkan keakuratan yang tinggi untuk beberapa jenis koleksi, pemilihan dari fungsi kernel tidak jelas dan dapat bergantung pada jenis koleksi. Dalam hasil eksperimen, nilai ketepatan tertinggi dihasilkan oleh fungsi kernel Gaussian RBF antara histogram label simpul (KV_{Gauss}).

Dalam penelitian ini digunakan klasifikasi SVM, menurut Kim et all (2005) SVM telah diakui sebagai salah satu metode klasifikasi untuk berbagai aplikasi termasuk klasifikasi teks. Nilai ketepatan prediksi dengan penggunaan SVM lebih baik karena kesalahan generalisasi yang rendah dengan memaksimalkan margin, dan kemampuan untuk menangani non-linearitas dengan pilihan kernel (Kim, Howland, & Park, 2005).

PENUTUP

Artikel ilmiah dalam suatu pangkalan data perpustakaan digital diklasifikasikan pada kategori-kategori tertentu. Pemodelan data artikel ilmiah menjadi graph bipartite dapat mengklasifikasi artikel ilmiah dengan menggunkan metode berbasis kernel. Dalam penelitian ini diperkenalkan lima macam kernel graph, yaitu: kernel Gaussian RBF antara histogram label sisi (KE_{Gauss}), kernel linier antara histogram label sisi (KE_{Linear}), kernel Gaussian RBF antara histogram label simpul (KV_{Gauss}), kernel linier antara histogram label simpul (KV_{Linear}) dan kernel Random Walk (*KRW*). Teknik matriks kernel dapat digunakan sebagai input untuk pengklasifikasi SVM dalam pengkategorian artikel ilmiah dalam pangkalan data perpustakaan digital secara otomatis. Penerapan metode ini dalam pangkalan data ISJD (*Indonesian Scientific Journal Database*) menghasilkan rata-rata ketepatan yang signifikan yaitu 87,43% untuk fungsi kernel KV_{Gauss}. Sedangkan kernel lainnya memberikan hasil berturut-turut 86,14% (KE_{Linear}), 85,86% (KE_{Gauss}), 42,23% (KV_{Linear} dan 25,15% (*KRW*).

Hasil ini menunjukkan bahwa penggunaan metode kernel graf efektif untuk mengelompokkan artikel ilmiah ke dalam kategori yang ditentukan dalam pangkalan data perpustakaan digital.

DAFTAR PUSTAKA

Andrej, M., and Doreian, P.”Partitioning signed two-mode networks”. *Journal of Mathematical Sociology,* 33(2009): 196–221

Banerjee, S., K. Sarkar, S. Gokalp, A. Sen, and H. Davulcu. “Partitioning signed bipartite graphs for classification of individuals and organizations”. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7227 LNCS*, 196–204. 2012.

Dhillon, I. S. “Co-clustering documents and words using Bipartite Spectral Graph Partitioning”. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’01*, San Francisco, CA, USA,(2001) pp. 269–274.

de Paulo Faleiros,T., Rossi, R.G., and de Andrade Lopes,A. “Optimizing the class information divergence for transductive classification of texts using propagation in bipartite graphs”. *Pattern Recognition Letters*. (2016). http://dx.doi.org/10.1016/j.patrec.2016.04.006

Grace, G.H. and Desikan, K. “Document clustering using a new similarity measure based on energy of a bipartite graph”. *Indian Journal of Science and Technology* 9(40) (2010). http://dx.doi.org/10.17485/ijst/2016/v9i40/99005

Kim, H., Howland, P., & Park, H. Dimension Reduction in Text Classification with Support Vector Machines. *Journal ofMachine Learning Research*, *6* (2005): 37–53. https://doi.org/10.1021/bi702018v

Martın De Diego, I., A. Munoz, and J. M. Moguerza. “Methods for the combination of kernel˜ matrices within a support vector framework”. *Machine Learning 78*(1-2) (2010): 137–174.

Radev, D. R. “Weakly supervised graph-based methods for classification”. *Ann Arbor 1001*(1) (2009): 48109–1092.

Srivastava, A., A. Soto, and E. Milios. “Text clustering using one-mode projection of documentword bipartite graphs”. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing - SAC ’13*, Coimbra, Portugal, (2013): 927–932.

Stankova, M., D. Martens, and F. Provost. “Classification over Bipartite Graphs through Projection”. Technical Report D/2015/1169/001, University of Antwerp, Antwerp, Belgium Research. 2015.

Sugiyama, M. and K. Borgwardt. “Halting in Random Walk Kernels”. *Advances in Neural Information Processing Systems* (Section 2) (2015): 1639–1647.

Yoo, I., X. Hu, and I.-y. Song. “Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering”. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’06*, Philadelphia, PA, USA, (2006): 791.

Zha, H., X. He, C. Ding, M. Gu, and H. Simon. “Bipartite graph partitioning and data clustering”. In *Proceedings of the tenth international conference on Information and knowledge management - CIKM ’01*, Volume pages, Atlanta, Georgia, USA, (2001): 25.

Zha, H. and X. Ji. “Correlating multilingual documents via bipartite graph modeling”. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR ’02*, Tampere, Finland, (2002): 443.