Ekstraksi Ciri Produktivitas Dinamis untuk Prediksi Topik Pakar dengan Model *Discrete Choice*

Diana Purwitasari[1,2], Chastine Fatichah[2], Surya Sumpeno[1,3], Mauridhi Hery Purnomo[1,3]

***Abstract*—Recommendation of active or productive experts is indispensable in supporting collaborations. Activities of publication and citation indicate expert productivity. An expert can be inferred to have an interest in a subject through productivity in that particular topic. Since an expert can change interests over time, the contribution of this paper is a Discrete Choice Model (DCM) based on topic productivities to predict the primary interests of the experts. DCM uses features extracted from bibliographic data of citation relation and title-abstract texts. Before extracting productivity features and dynamicity features to represent interest changes, title clustering with KMeans++ is used to identify research topics. There are six productivity features and five dynamicity values for each productivity feature to demonstrate the expert behavior. Therefore, a clustered topic as a research interest is represented as an expert choice with 30 extracted features in the proposed method. The experiments used multinomial logistic regression for DCM and a log-likelihood indicator for the fitted models of the features. The resulted DCM models showed that productive behavior of the experts by doing many publications and receiving many citations effected to the precision of topic prediction by 80%. Some features were better for predicting primary interests of the expert. It was demonstrated with a lower precision value of 60% by using features that represent the expert behavior of only doing publication or only getting citation.**

***Intisari*—Rekomendasi pakar yang aktif atau produktif dapat mendukung kelancaran kolaborasi penelitian. Kegiatan publikasi dan sitasi menjadi indikasi produktivitas seorang pakar. Konfirmasi ketertarikan pakar akan suatu subjek terlihat dengan produktivitasnya di topik tersebut. Dikarenakan fokus riset pakar dapat berubah, kontribusi pada makalah ini adalah model *Discrete Choice* (*Discrete Choice Model*, DCM) untuk prediksi topik pakar yang dianggap utama berdasarkan produktivitasnya. DCM menggunakan ciri hasil ekstraksi data bibliografi, yaitu relasi sitasi dan teks judul serta intisari dari artikel milik pakar. Sebelum proses ekstraksi ciri produktivitas dan ciri dinamis untuk representasi perubahan fokus riset dari tiap pakar, pengklasteran teks judul dengan *KMeans++* dilakukan untuk identifikasi topik riset. Ada enam jenis ciri produktivitas dan lima kriteria dinamis pada tiap ciri untuk merepresentasikan perilaku pakar terhadap fokus risetnya, sehingga untuk setiap topik hasil pengklasteran akan direpresentasikan sebagai suatu pilihan bidang riset dari seorang pakar dengan 30 ciri hasil ekstraksi. Uji coba yang dilakukan menggunakan pendekatan regresi logistik multinomial untuk DCM dan *log-likelihood* sebagai indikator pengujian model dari ciri. DCM yang dihasilkan menunjukkan perilaku pakar terkait produktivitas dengan aktif publikasi dan banyak menerima sitasi memberikan nilai presisi pada proses prediksi topik sebesar 80%. Beberapa kombinasi ciri memberikan hasil prediksi yang lebih baik. Hal tersebut terlihat dengan nilai presisi yang lebih rendah, yaitu 60%, jika prediksi topik dilakukan dengan model dari ciri terkait banyak publikasi saja atau banyak sitasi saja.**

***Kata Kunci*—prediksi topik, ekstraksi ciri produktivitas, profil pakar, model *discrete choice*, data bibliografi.**

I. PENDAHULUAN

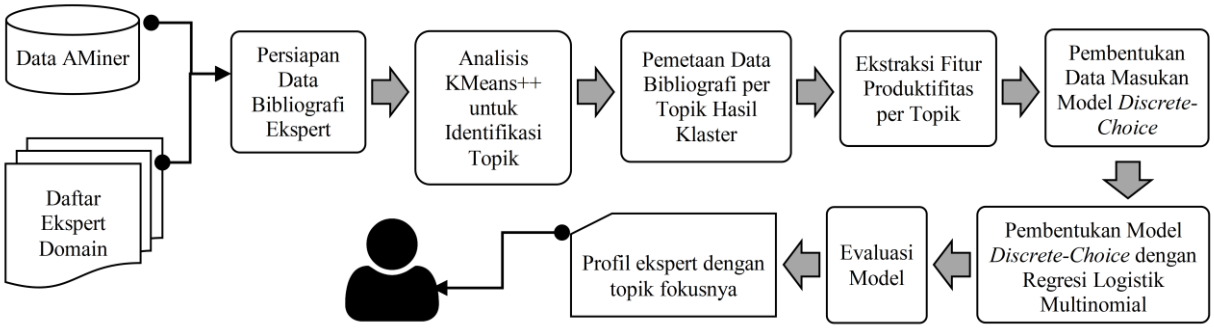
Pencarian pakar, disebut juga *expert finder* dalam sistem rekomendasi, memberikan keluaran daftar pakar berdasarkan pemeringkatan [1] yang mendukung kolaborasi pakar sebagai konsultan [2] maupun pembimbing [3] dalam kegiatan bersama. Data masukan adalah bibliografi publikasi ilmiah [4], seperti judul, pakar sebagai penulis, referensi, sitasi, serta metadata lain. Sistem memodelkan pakar, dokumen hasil karyanya, serta topik yang diekstrak [5]. Pemodelan berbasis peluang, *voting*, representasi graf, atau kombinasi memberikan informasi pakar berpengalaman secara kuantitatif dan kualitatif di suatu topik riset. Pendekatan berbasis peluang membuat peringkat dengan menghitung kemungkinan pakar disebut ahli dalam topik tertentu [6]. Keahlian pakar dilihat dari artikel ilmiah yang dipublikasikannya. Keaktifan atau produktivitas pakar dapat ditunjukkan dengan jumlah artikel yang dipublikasikan serta sitasi yang didapat di suatu periode pengamatan [7]. Metode pemeringkatan secara *voting* [7] mengembalikan dokumen pakar berdasarkan produktivitas pakar dengan jumlah artikel serta sitasi di suatu periode pengamatan. Pada model graf, *node* dominan dicari sebagai pakar yang memiliki keahlian [8].

Pakar dapat memiliki banyak fokus topik penelitian, antara lain sebagai hasil kolaborasi antar domain [9]. Penelitian yang telah ada membuat pemodelan pakar dengan graf, tetapi tidak mengakomodasi topik penelitian [2]. Pemodelan topik berbasis peluang tidak memperlihatkan informasi produktivitas atau keaktifan pakar [6]. Pada pemodelan banyak topik, suatu topik memiliki representasi graf yang digunakan untuk menghitung skor keahlian pakar [10]. Pakar dengan skor tinggi di suatu bidang belum tentu masih aktif meneliti atau produktif di topik tersebut. Oleh karena itu, pemodelan pakar yang sudah ada belum mempertimbangkan keaktifan atau produktivitas pakar dalam menentukan skor keahlian pakar. Isu tersebut yang menjadi fokus kontribusi dalam makalah ini.

[1]*Departemen Teknik Elektro, Fakultas Teknologi Elektro (FTE), Institut Teknologi Sepuluh Nopember (ITS) Kampus ITS, Sukolilo, Surabaya, Jawa Timur, 60111 Indonesia (tlp: 031- 5939214; fax: 031-5913804; email: diana@if.its.ac.id; diana.purwitasari@gmail.com)*

[2]*Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi (FTIK), Jln. Teknik Kimia, Gedung Informatika, Institut Teknologi Sepuluh Nopember (ITS) (Telp: +62-31-5939212, 5995581, 5994251 ext 1445, 1446; email: chastine@if.its.ac.id)*

[3]*Departemen Teknik Komputer, Fakultas Teknologi Elektro (FTE), Institut Teknologi Sepuluh Nopember (ITS) Gedung B & C, Kampus ITS Sukolilo Surabaya (tlp: 031-5994251-54, 5947274, 5945472; email: surya@ee.its.ac.id; hery@ee.its.ac.id)*



Gbr. 1 Tahapan pemodelan pakar dari data bibliografi dengan pendekatan model *Discrete Choice.*

Produktivitas memiliki bobot kuantitatif dan kualitatif berbeda yang dipengaruhi oleh waktu [11]. Ciri produktivitas telah diteliti untuk mengelompokkan pakar berdasarkan potensinya. Sebagai ilustrasi, terdapat dua pakar aktif atau produktif dengan jumlah publikasi dan sitasi yang sama di suatu topik. Namun, pakar aktif saat ini seharusnya memiliki skor keahlian lebih tinggi dibanding pakar dengan masa aktif beberapa tahun sebelumnya karena produktivitas dipengaruhi kuantitas dan kualitas. Kuantitas adalah jumlah publikasi atau sitasi suatu pakar di suatu waktu pengamatan, sedangkan kualitas atau kontinuitas adalah jumlah publikasi atau sitasi di suatu topik secara akumulatif sampai saat pengamatan.

Kontribusi makalah ini adalah pemodelan dengan ekstraksi ciri produktivitas pakar untuk prediksi topik dominan. Banyaknya topik diasumsikan sebagai banyaknya pilihan. Berbeda dengan pemodelan bibliografi berbasis peluang kemunculan teks untuk topik, *voting* untuk peringkat dokumen atau representasi dokumen dan atau pakar dalam graf, DCM menghitung kualitas alternatif pilihan. Penggunaan DCM memberikan perbaikan adanya sudut pandang produktivitas yang berubah dipengaruhi waktu atau dinamis di setiap fokus topik riset pakar jika dibanding pemodelan graf [2] atau peluang [6] yang statis.

DCM sering digunakan dalam prediksi [12] atau rekomendasi [13]. DCM dalam permasalahan prediksi menghitung peluang kualitas suatu tampilan laman web dibanding laman lainnya. Sedangkan pada rekomendasi, DCM menghitung kualitas peluang pilihan setiap *item*. DCM dengan pendekatan model estimasi dapat menjelaskan perilaku subjek terhadap pilihan yang memengaruhi keputusan akhir [14]. Hal tersebut tidak dapat ditunjukkan oleh pembelajaran mesin dengan pemodelan pakar berbasis peluang, *voting*, atau graf, sehingga DCM lebih sesuai untuk prediksi topik dominan milik pakar karena dapat menjelaskan perilaku ketertarikan pakar pada topik tersebut.

Struktur bahasan dalam makalah ini adalah sebagai berikut. Pada bagian II diuraikan usulan metode berbasis DCM untuk representasi bibliografi. Bahasan diawali dengan deskripsi data bibliografi, ekstraksi ciri produktivitas dari data bibliografi, dan pemodelan regresi logistik multinomial dengan ciri-ciri. Bagian III membahas uji coba mulai dari analisis pengklasteran topik sampai pencarian model yang paling sesuai dengan variasi ciri produktivitas.

II. USULAN METODE UNTUK REPRESENTASI BIBLIOGRAFI

Tahapan usulan metode untuk pemodelan pakar dari data bibliografi ditunjukkan pada Gbr. 1, mulai dari persiapan data sampai pembentukan model *Discrete Choice*.

A. Deskripsi Data Bibliografi

Data bibliografi adalah data-data terkait suatu artikel ilmiah [4]. Ada banyak penelitian terkait dengan pengolahan data bibliografi menjadi informasi. Secara detail, metadata artikel ilmiah terdiri atas judul artikel, intisari, kata kunci, keterangan penulis termasuk asal instansi, daftar referensi atau disebut juga sitasi. Metadata artikel dari asal datanya dikategorikan sebagai ciri konten dan ciri struktur. Ciri berdasarkan konten adalah teks dari judul, intisari, dan semacamnya. Kemunculan kata adalah nilai yang sering dijadikan ciri. Namun, banyaknya jumlah kata memengaruhi dimensi ciri. Kemudian, ciri berdasarkan struktur adalah semua relasi yang dapat diambil, antara lain antar penulis, antar artikel dalam bentuk sitasi, atau relasi antar topik artikel. Penelitian dengan data bibliografi sering mengombinasikan kedua jenis dasar ciri. Pada makalah ini digunakan teks judul dan intisari yang diolah menjadi topik untuk mengurangi dimensi ciri konten. Sedangkan ciri struktur diwakili dengan sitasi. Koleksi data bibliografi yang digunakan diuraikan pada bagian uji coba.

B. Ekstraksi Fitur Produktivitas

Penelitian dalam makalah ini menggunakan data bibliografi yang sudah ada tetapi masih memerlukan tahap persiapan data untuk domain tertentu (*sampling* data). Tidak ada informasi topik yang menjadi label di setiap pakar dan dokumen hasil publikasinya, sehingga dilakukan pengklasteran *KMeans++* dan ekstraksi label untuk analisis hasil klaster [15] pada Gbr. 1. Detail proses pengklasteran, penentuan jumlah topik (nilai K), dan analisisnya diuraikan pada bagian uji coba.

Klaster hasil adalah klaster topik c_i beranggotakan kata-kata terkait. Tidak semua kata dalam teks koleksi bibliografi yang masuk dalam klaster topik. Pada suatu dokumen d_a, kata-kata penting t_b diekstrak dan jika suatu kata masuk dalam klaster, $t_b \in c_i$, maka dokumen yang berisi kata tersebut dilabeli dengan klaster yang sama, $d_a \in c_i$. Setiap kata hanya masuk dalam satu klaster topik, tetapi suatu dokumen dapat dipetakan ke lebih dari satu topik, sehingga pakar a_x yang menjadi penulis dokumen juga diasumsikan tertarik pada topik tersebut.

TABEL I
FITUR PRODUKTIVITAS DIMILIKI OLEH PAKAR PER TOPIK

Nomor Ciri	Deskripsi Ciri	Keterangan Pengamatan
$F_1(a_x, c_i, t_n)$	Jumlah dokumen	pada periode pengujian t_n
$F_2(a_x, c_i, t_m, t_n)$	Jumlah dokumen secara kumulatif (1)	dari t_m sampai periode pengujian t_n
$F_3(a_x, c_i, t_m, t_n)$	Bobot jumlah dokumen secara kumulatif (2)	dari t_m sampai periode pengujian t_n
$F_4(a_x, c_i, t_n)$	Jumlah sitasi (3)	pada periode pengujian t_n
$F_5(a_x, c_i, t_m, t_n)$	Jumlah sitasi secara kumulatif (4)	dari t_m sampai periode pengujian t_n
$F_6(a_x, c_i, t_m, t_n)$	Bobot jumlah sitasi secara kumulatif (5)	dari t_m sampai periode pengujian t_n

Produktivitas pakar ditandai ekstraksi nilai di suatu periode pengamatan atau akumulasinya dari unsur publikasi dan unsur sitasi [11]. Namun, belum ada informasi topik dalam ekstraksi ciri yang sudah dilakukan, sehingga pada makalah ini dilakukan modifikasi ekstraksi ciri produktivitas yang ditunjukkan di Tabel I. Ciri-ciri tersebut diperhitungkan dari koleksi dokumen yang dipublikasikan oleh pakar a_x. Ciri F_1, F_2, F_3 terkait jumlah publikasi dan ciri F_4, F_5, F_6 terkait jumlah sitasi.

$$F_2(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} F_1(a_x, c_i, t_y) \quad (1)$$

$$F_3(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} \frac{F_1(a_x, c_i, t_y)}{t_y - t_m + 1} \quad (2)$$

$$F_4(a_x, c_i, t_n) = \sum_{d_a \in c_i} ncite(d_a, t_n) \quad (3)$$

$$F_5(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} F_4(a_x, c_i, t_y) \quad (4)$$

$$F_6(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} \frac{F_4(a_x, c_i, t_y)}{t_y - t_m + 1} \quad (5)$$

Ciri F_1 adalah jumlah dokumen pada suatu klaster topik $d_a \in c_i$ di tahun pengamatan t_n. Ciri F_2 adalah akumulasi jumlah publikasi F_1 sedangkan F_3 adalah nilai F_2 yang disertai bobot penalti sebagai penanda kontinuitas ketertarikan pakar di suatu topik. Untuk ciri F_4, fungsi $ncite(d_a, t_n)$ adalah jumlah sitasi yang diperoleh d_a di tahun pengamatan t_n dengan catatan dokumen tersebut masuk dalam klaster topik $d_a \in c_i$.

Aspek dinamis di setiap ciri produktivitas (misalnya F_1) pada suatu topik tertentu untuk menangkap kinerja pakar adalah perubahan mínimum (*min*), maksimum (*max*), terakhir (*end*), serta total perubahan (*sum*) di (6)-(9). Kemudian nilai representatif kinerja keseluruhan pakar (*rep*) untuk F1 dan F4 dihitung dengan (10), untuk F2 dan F5 dengan (11), serta untuk F3 dan F6 dengan (12). Sebagai catatan, urutan waktu yang digunakan dari terkecil adalah $t_m, \dots t_{y-1}, t_y \dots t_{n-2}, t_{n-1}, t_n$.

$$F_{1.min}(a_x, c_i, t_m, t_n) = \min_{t_m \dots t_n} F_1\left(a_x, c_i, t_{y-1}, t_y\right) \quad (6)$$

$$F_{1.max}(a_x, c_i, t_m, t_n) = \max_{t_m \dots t_n} F_1\left(a_x, c_i, t_{y-1}, t_y\right) \quad (7)$$

$$F_{1.end}(a_x, c_i, t_{n-2}, t_n) = F_1(a_x, c_i, t_{n-1}, t_n) - F_1(a_x, c_i, t_{n-2}, t_{n-1}) \quad (8)$$

$$F_{1.sum}(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} F_1\left(a_x, c_i, t_{y-1}, t_y\right) \quad (9)$$

$$F_{1.rep}(a_x, c_i, t_m, t_n) = \sum_{t_m \dots t_n} \frac{F_1(a_x, c_i, t_y)}{t_n - t_m + 1} \quad (10)$$

$$F_{2.rep}(a_x, c_i, t_m, t_n) = \frac{F_2(a_x, c_i, t_{n-1}, t_n)}{t_n - t_m + 1} \quad (11)$$

$$F_{3.rep}(a_x, c_i, t_m, t_n) = F_3(a_x, c_i, t_{n-1}, t_n) \quad (12)$$

Ciri produktivitas dari pakar dihitung menggunakan (1)-(5). Setelah ekstraksi aspek dinamis dari ciri produktivitas dengan (6)-(12), kemudian dilakukan transformasi matriks ciri produktivitas menjadi model *Discrete Choice*.

C. Pemodelan Discrete Choice

DCM [16] merupakan teknik statistik yang sering digunakan dalam analisis pemasaran untuk evaluasi pilihan pelanggan terhadap produk atau layanan [17], seperti prediksi kualitas laman web [12], rekomendasi layanan hotel [13], atau layanan kesehatan [18], serta analisis kebiasaan pejalan kaki [19]. Kasus klasik DCM adalah pilihan transportasi perjalanan [20]. Jika suatu pilihan transportasi dipengaruhi waktu dan uang, maka model Regresi Linear terbentuk $y_{via.1} = \beta_0 + \beta_1 x_{wkt.1} + \beta_2 x_{rp.1} + \varepsilon$ memiliki variabel tidak bebas (*dependent variable*), yaitu nilai peluang pilihan $y_{via.1}$ dan variabel bebas (*independent variable*) yang memengaruhi keputusan pilihan seperti $x_{wkt.1}$ atau $x_{rp.1}$. Model Regresi Logistik digunakan untuk memastikan nilai peluang pilihan transportasi $y_{via.1}$ antara 0 dan 1. Model disebut Regresi Logistik Multinomial jika memiliki banyak pilihan. Topik riset yang diminati pakar ada batasannya, tetapi terdapat banyak kemungkinan kombinasi fokus penelitian.

DCM membutuhkan data latih sebagai masukan dengan tiga komponen pada Tabel II, yaitu individu, alternatif pilihan, serta pilihan individu. Fungsi utilitas U_l mengukur kualitas pilihan alt_l yang dipengaruhi fungsi variabel V_l. Implementasi komponen DCM pada permasalahan pakar ditunjukkan pada Tabel II.

TABEL II
KOMPONEN DASAR DCM

No	Komponen DCM	Keterangan	Implementasi komponen DCM
1	Individu	n	Pakar n
2	Alternatif Pilihan	alt_l dari himpunan J	Pakar n memiliki publikasi di Topik-l
3	Pilihan Individu	alt_l dipilih jika $U_l > U_j$ dengan $alt_l \neq alt_j$	$U_{ln} = V_{ln} + \varepsilon$

TABEL III
CONTOH HASIL EKSTRAKSI ENAM CIRI PRODUKTIVITAS DARI DATA BIBLIOGRAFI

ID	Author	Tahun_Uji	T0f1	T0f2	T0f3	T0f4	T0f5	T0f6
40	1139547	2009	0	3	0,282	4	38	11,217
41	1201431	2000	0	2	0,667	13	35	22,833
42	1201431	2001	0	2	0,500	15	50	28,250
43	1201431	2002	0	2	0,400	8	58	24,983
44	1201431	2003	1	3	1,333	16	74	32,417
45	1201431	2004	0	3	0,786	5	79	25,517

TABEL IV
CONTOH CIRI PRODUKTIVITAS YANG TELAH DIPROSES MENJADI ASPEK DINAMIS PAKAR

ID	Individual	Mode	choice	F1_min	F1_max	F1_end	F1_sum	F1_rep	F2_min	F2_max
9.topic28	1318540	Topic28	False	-0,33333	0,00000	0,00000	-0,33333	0,10000	0	0,00000
9.topic29	1318540	Topic29	True	-0,66667	1,00000	0,20000	0,33333	0,44444	0	0,66667
10.topic0	1351785	Topic0	False	-0,33333	0,33333	0,00000	-0,25000	0,09091	0	0,20000
10.topic1	1351785	Topic1	False	0,00000	0,00000	0,00000	0,00000	0,00000	0	0,00000
10.topic2	1351785	Topic2	False	-0,66667	0,50000	0,00000	-0,25000	0,10714	0	0,40000

Pada Tabel III dan Tabel IV terlihat dua pakar. Pada Tabel III hasil ekstraksi enam ciri produktivitas diambil dari data bibliografi untuk dua pakar dengan periode sepuluh tahun pengamatan (2000-2009) pada suatu topik (T0). Jumlah baris matriks adalah jumlah pakar x jumlah tahun pengamatan dan jumlah kolom matriks adalah jumlah ciri produktivitas x jumlah topik. Sedangkan contoh kolom [*mode*] di Tabel IV menunjukkan pilihan pakar-9 adalah topic29 yang artinya nilai $U_{l=29}$ lebih besar dibanding pilihan lain. Pada Tabel IV ciri produktivitas telah diproses menjadi aspek dinamis pakar sebagai masukan model *Discrete Choice*. Jumlah baris matriks adalah jumlah pakar x jumlah topik dan jumlah kolom matriks adalah jumlah ciri produktivitas x jumlah ciri dinamis. Fungsi penentuan pilihan U_l dengan nilai *True/False* sebagai *pilihan* pada kolom [*choice*] diuraikan kemudian.

Alternatif pilihan adalah topik yang menjadi fokus penelitian dengan variabel-variabel bebasnya V_l adalah nilai hasil ekstraksi ciri-ciri produktivitas. Pada Tabel IV kolom [*mode*] terdapat 30 alternatif pilihan, topic0 sampai topic29. Ketertarikan pakar terhadap suatu topik direpresen-tasikan dengan enam ciri produktivitas yang telah diproses kinerja dinamisnya, yaitu:

a. nilai perubahan mínimum $F_{1.min} \dots F_{6.min}$,
b. nilai perubahan maksimum $F_{1.max} \dots F_{6.max}$,
c. nilai perubahan terakhir $F_{1.end} \dots F_{6.end}$,
d. nilai total perubahan $F_{1.sum} \dots F_{6.sum}$, dan
e. nilai representatif kinerja keseluruhan $F_{1.rep} \dots F_{6.rep}$.

Pemodelan pakar dengan DCM adalah sebagai berikut. Pakar a_x atau individu n sebagai pengambil keputusan memiliki topik yang diminati, disebut alternatif j dari himpunan pilihan J. Atribut atau variabel z_{jn} yang mendeskripsikan pilihan adalah 30 ciri produktivitas setelah diproses kinerja dinamisnya. Fungsi utilitas U_{jn} dapat menunjukkan peluang individu n memilih alternatif topik j sebagai fokus risetnya dengan model pada (13). Pada tahap uji coba diuraikan penentuan nilai ambang $thres$ yang menjadi batasan fungsi utilitas U_{jn} sesuai model pada (14).

$$U_{jn} = \beta_{j1.min}F_{j1.min}(n) + \cdots + \beta_{j1.rep}F_{j1.rep}(n) + \cdots + \beta_{j6.min}F_{j6.min}(n) + \cdots + \beta_{j6.rep}F_{j6.rep}(n) + \varepsilon_{jn} \quad (13)$$

$$P(j|J_n) = P\left[U_{jn} > thres\right]_{j \in J_n} \quad (14)$$

Analisis kombinasi nilai ciri diuraikan pada skenario uji coba untuk mengetahui perilaku pakar yang memengaruhi pilihan topik, yaitu dalam hal menghasilkan publikasi atau mendapatkan sitasi.

III. HASIL DAN ANALISIS

A. Dataset

Dataset terdiri atas koleksi publikasi ilmiah dengan identitas pemakalah dan metadata artikel, antara lain judul, abstrak, serta kata kunci. Penelitian terkait pakar dan bibliografi sering memanfaatkan *dataset Association for Computational Linguistics* (ACL) [21] atau AMiner [8] pada tahap uji coba, meskipun banyak terdapat ketidaklengkapan metadata. Pada makalah ini digunakan daftar pakar di beberapa domain topik penelitian tertentu[1], yaitu *Natural Language Processing* (54 pakar di bidang NLP) dan *Information Extraction* (91 pakar di bidang IE) yang diistilahkan sebagai Daftar Pakar Domain. Sebagai catatan, domain NLP dan IE memiliki keterkaitan bahasan. Koleksi publikasi ilmiah AMiner dalam himpunan data D_{big} memiliki

a. daftar publikasi ilmiah beserta kode artikel (± 1.9M artikel),
b. daftar pemakalah beserta kode penulis (± 1.6M pakar), dan
c. daftar kolaborasi antar penulis (± 4.2M relasi).

[1] https://aminer.org/data#Expert-Finding

TABEL V
NILAI *SILHOUETTE* SEBAGAI INDIKATOR PENENTUAN JUMLAH KLASTER TOPIK

Dataset	**Min DF**	**Jumlah Klaster**	**No Uji**	**Avg Silh**	**Algoritme**	**Matriks**	**Jenis Ciri**	**Deskripsi**
D_{small} (±14.000 data x 200 dimensi)	10	100	1	0,028	*KMeans++*	± 3.500 x 200	DF	*Document Frequency*
	3	100	2	0,005	*KMeans++*	± 7.000 x 200		
	3	50	3	0,068	*KMeans++*	± 7.000 x 200		
	10	100	4	-0,022	*KMeans++*	± 3.500 x 200		
	10	100	5	**0,264**	*GaussMix*	± 3.500 x 2	FeatAgglo(2)	*Feature Agglomeration*
	10	100	6	0,128	*GaussMix*	± 3.500 x 2	PCA(2)	*Principal Component Analysis*
	10	50	7	-0,028	*KMeans++*	± 3.500 x 200		
D_{big} (±62.500 data x 200 dimensi)	10	100	8	-0,046	*KMeans++*	± 12.000 x 200		
D_{small} (±14.000 data x 100 dimensi)	10	100	9	0,135	*KMeans++*	± 3.500 x 100		
			10	**0,364**	*GaussMix*	± 3.500 x 2	FeatAgglo(2)	*Gaussian Mixture*
			11	0,003	*GaussMix*	± 3.500 x 10	FeatAgglo(10)	
			12	0,256	*GaussMix*	± 3.500 x 2	PCA(2)	
			13	0,104	*GaussMix*	± 3.500 x 10	PCA(10)	
		50	14	0,115	*KMeans++*	± 3.500 x 100		
		30	15	**0,179**	*KMeans++*	± 3.500 x 100		
			16	**0,177**	*GaussMix*	± 3.500 x 2	FeatAgglo(2)	
			17	0,158	*GaussMix*	± 3.500 x 10	FeatAgglo(10)	
$D_{small-title}$ (±4.200 data x 100 dimensi)	10	30	18	0,215	*KMeans++*	± 600 x 100		
			19	0,527	*GaussMix*	± 600 x 2	FeatAgglo(2)	
			20	**0,651**	*GaussMix*	± 600 x 2	FeatAgglo(2)	
			21	0,206	*GaussMix*	± 600 x 2	PCA(2)	

Namun, belum ada relasi antara artikel ilmiah dan pemakalah yang menulis artikel tersebut, sehingga pemrosesan semi manual dilakukan untuk mendapatkan artikel milik pakar yang terdapat dalam Daftar Pakar Domain. Beberapa pakar memiliki kemungkinan nama ambigu atau jumlah artikel dalam koleksi tidak mencukupi (< 20 artikel). Setelah validasi data, maka *dataset* untuk uji coba terdiri atas 70 pakar dan ±4,8K artikel yang disebut himpunan data D_{small}. Koleksi data tersebut hanya terdiri atas ±75% artikel yang memiliki teks abstrak karena metadata tidak lengkap.

Jumlah kata terindeks pada D_{small} adalah ±14K kata dengan ukuran matriks *word embedding* ±3,5K x 100 dimensi untuk *min_df* =10. Jumlah paper yang disitasi (*cited paper*) adalah 2.366 (0.52%) dan jumlah paper yang melakukan sitasi (*citation*) adalah 7,771 (1.69x)

B. Skenario Uji Coba

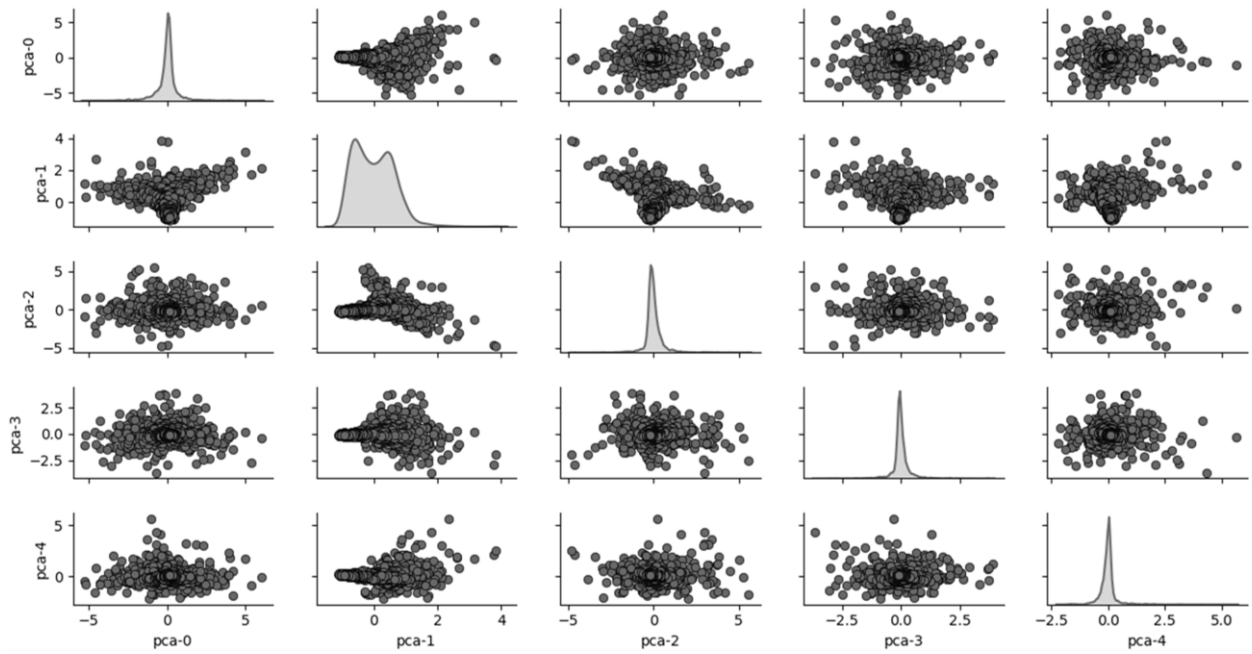
Ada beberapa hal yang dilakukan sebagai bagian dari uji coba pemodelan pakar dan fokus risetnya menjadi bentuk DCM.

1. Menentukan jumlah topik yang digunakan sebagai fokus riset karena *dataset* tidak menyediakan informasi tersebut. Uji coba menggunakan beberapa pendekatan pengklasteran serta metode ekstraksi ciri yang berbeda (Tabel V). Analisis ekstraksi fitur dilakukan untuk mendeskripsikan hasil klaster topik (Gbr. 2 - Gbr. 4).
2. Kemudian setiap pakar dipetakan ke topik yang dianggap fokus risetnya sebagai pembentukan *ground-truth* karena informasi tersebut tidak tersedia (Gbr. 5). Data *ground-truth* digunakan pada validasi model hasil DCM.
3. Pada pembuatan data *ground-truth* dilakukan analisis nilai ambang fungsi utilitas DCM (Gbr. 6) untuk menentukan topik yang dianggap fokus riset utama pakar.
4. Menentukan model dan kombinasi ciri yang paling sesuai untuk melakukan prediksi fokus riset utama dari pakar (Tabel VI dan Tabel VII). Perbandingan dengan algoritme klasifikasi lainnya juga ditunjukkan di Tabel VIII.

C. Hasil dan Analisis Proses Identifikasi Topik

Data teks dengan kata sebagai ciri mempunyai masalah klasik terkait jumlah dimensi, sehingga pendekatan Word2Vec dipilih untuk mengurangi kompleksitas waktu [22]-[24].

Proses *word embedding* pada koleksi (±4,8K artikel) dengan Python Library Gensim [25] mengabaikan kata t yang muncul di sepuluh dokumen (*document frequency* $df_t > 10$ atau 0,002) dan jumlah dimensi = 200. Hasil akhir adalah ±3.500 kata (±25%) dari ±14.000 kata terindeks berupa matriks *sparse* berukuran ±3.500 x 200. Elemen matriks dapat bernilai negatif yang berarti kata memiliki sentimen berlawanan pada dimensi tersebut.



Gbr. 2 *Plot pairwise* dari fitur kata D_{small} dalam klaster setelah transformasi dua dimensi (2D) dengan PCA 5 komponen berisi teks judul dan abstrak. *Plot* kurva di arah diagonal adalah distribusi nilai dan densitas setiap komponen PCA.

K-Means++ menggunakan Python Library scikit-learn [26] pada hasil *word embedding* untuk inisialisasi titik tengah klaster (*cluster centroid*) mencari kelompok topik dengan anggota berupa kata yang sering digunakan untuk membahas topik tersebut. Jumlah klaster ditentukan hingga dalam satu klaster terdapat ±1,5% data dari total koleksi data (*num_cluster* = 100).

Berbagai kombinasi dari himpunan data teks, algoritme pengklasteran, jumlah dimensi, dan reduksi dimensi fitur untuk analisis klaster dilakukan dengan hasil terlihat pada Tabel V.

Untuk setiap jenis himpunan data, nilai *silhouette* [27] terbaik atau lebih besar sebagai indikator hasil klaster ditunjukkan pada Uji 5, 10, 15, 16, dan 20. Klaster topik yang lebih baik terbentuk dari kata-kata di bagian judul artikel, yaitu himpunan data $D_{small-title}$ di Tabel V dengan dimensi *word embedding* = 100, sehingga untuk koleksi dokumen uji coba dipilih 30 topik sebagai alternatif (himpunan $J = \{j_1 \dots j_{30}\}$).

Bahasan dalam domain NLP dan IE terkait erat sehingga pengklasteran kata dari teks koleksi publikasi memiliki kemungkinan tumpang tindih (*overlap*). Sebagai catatan, terdapat ±10% data teks dalam domain IE yang memiliki kata-kata "*natural language processing*". Oleh karena itu, pengklasteran matriks hasil *word embedding* memiliki nilai *silhouette* yang rendah.

Analisis visual tentang kualitas klaster dan kata-kata sebagai fitur penyusun ditunjukkan pada Gbr. 2, Gbr. 3, dan Gbr. 4. Pada Gbr. 2 terlihat *plot pairwise* dari ciri yang telah ditransformasi dengan algoritme reduksi dimensi *Principal Component Analysis* (PCA) dari seratus ciri menjadi lima ciri (kode pca-0 sampai pca-4).

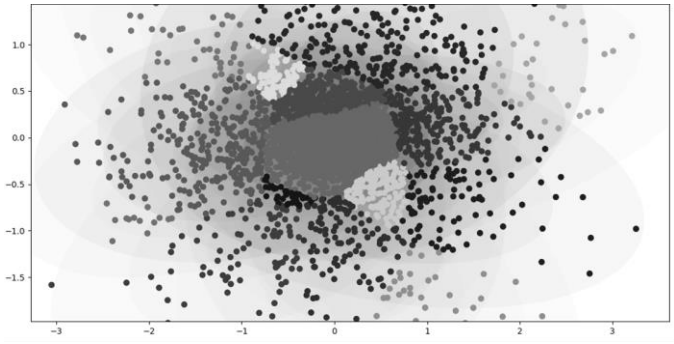
Plot kurva di arah diagonal menunjukkan distribusi nilai dan densitas setiap ciri PCA. Hanya nilai pca-1 yang menunjukkan sebaran lebih besar sehingga pengklasteran judul berdasarkan ciri tersebut menghasilkan kelompok yang cenderung terpisah. Terlihat bahwa nilai setiap ciri cenderung berkisar di nilai 0 sehingga plot sebaran nilai pasangan ciri PCA berbeda indeks akan berkumpul di titik 0 atau terlihat tumpang tindih.

Hal itu berakibat pada nilai *silhouette* yang rendah karena data tidak mudah untuk dipisahkan ke dalam kelompok berbeda seperti ditunjukkan pada Gbr. 2. Visual kata-kata untuk himpunan data yang sama dalam dua dimensi terlihat pada Gbr. 3. Namun, himpunan data berbeda yaitu $D_{small-title}$ berisi teks judul memberikan hasil pengklasteran lebih baik dengan analisis visual terlihat pada Gbr. 4. Hal tersebut terjadi karena kata-kata dalam teks judul cenderung ringkas dan pendek.

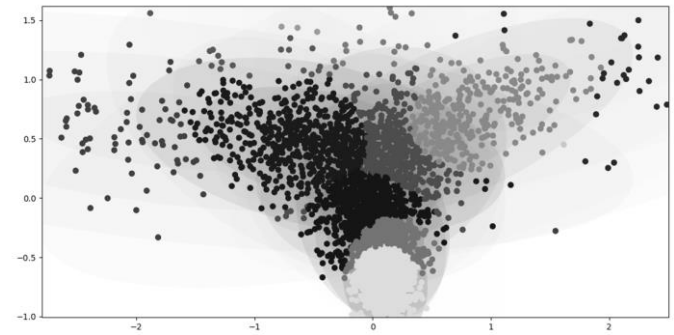
Setelah pemetaan topik ke setiap pakar, terlihat pada Gbr. 5 bahwa banyak pakar memiliki lebih dari satu fokus riset. Penelitian di domain NLP lebih dahulu berkembang karena kode pakar dalam *dataset* cenderung lebih kecil dibanding pakar domain IE.

D. Penentuan Nilai Ambang Fungsi Utilitas DCM

Hasil klaster dengan 30 topik dipetakan ke setiap dokumen dan pakar. Sebagian besar pakar pada *dataset* uji coba memiliki publikasi dari 10-20 topik berbeda. Validasi manual dilakukan pada tiap pakar untuk evaluasi kecenderungan fokus topik penelitian.



Gbr. 3 Plot kata-kata judul dalam klaster hasil *Bayesian Gaussian Mixture* yang telah ditransformasi 2D dengan PCA 2 komponen untuk D_{small}. Sumbu-x dan Sumbu-y adalah hasil transformasi nilai 100 fitur judul menjadi nilai pca-0 dan pca-1 dari Gbr. 2.



Gbr. 4 Plot kata-kata judul dalam klaster hasil *Bayesian Gaussian Mixture* yang telah ditransformasi 2D dengan PCA 2 komponen untuk $D_{small-title}$.Sumbu-x dan Sumbu-y adalah hasil transformasi nilai 100 fitur judul menjadi nilai pca-0 dan pca-1 dari Gbr. 2.

TABEL VI
NILAI *LOG.LIKELIHOOD* SEBAGAI INDIKATOR MODEL DCM

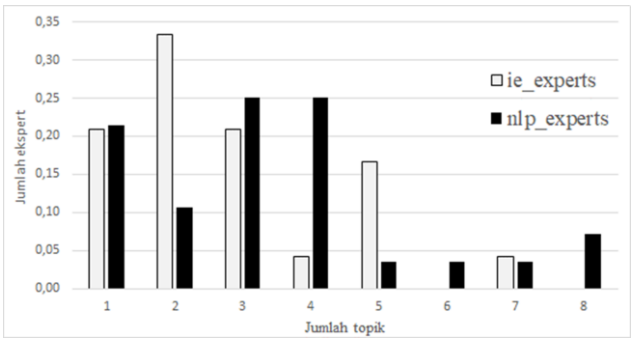
m	**Atribut Fungsi Utilitas**	***Log. likelihood***
1	Persamaan (13)	-0,0718
2	$\beta_{j1.min}F_{j1.min}(n) + \cdots + \beta_{j1.rep}F_{j1.rep}(n) + \varepsilon_{jn}$	-56,371
3	$\beta_{j2.min}F_{j2.min}(n) + \cdots + \beta_{j2.rep}F_{j2.rep}(n) + \varepsilon_{jn}$	-42,060
4	$\beta_{j3.min}F_{j3.min}(n) + \cdots + \beta_{j3.rep}F_{j3.rep}(n) + \varepsilon_{jn}$	-61,145
5	$\beta_{j4.min}F_{j4.min}(n) + \cdots + \beta_{j4.rep}F_{j4.rep}(n) + \varepsilon_{jn}$	-22,804
6	$\beta_{j5.min}F_{j5.min}(n) + \cdots + \beta_{j5.rep}F_{j5.rep}(n) + \varepsilon_{jn}$	-24,423
7	$\beta_{j6.min}F_{j6.min}(n) + \cdots + \beta_{j6.rep}F_{j6.rep}(n) + \varepsilon_{jn}$	-17,586
8	Semua nilai kinerja dinamis fitur $F_2 + F_6$	-3,66e-15

Berdasarkan sepuluh tahun pengamatan, dari 2000 hingga 2009, untuk ekstraksi ciri produktivitas dengan analisis kinerja dinamis, maka nilai $thres$ ditentukan dari tiap pakar dengan $\sum F_{j1.min} + \cdots + F_{j1.rep} + \cdots + F_{j6.min} + \cdots + F_{j6.rep} > 5{,}0$.

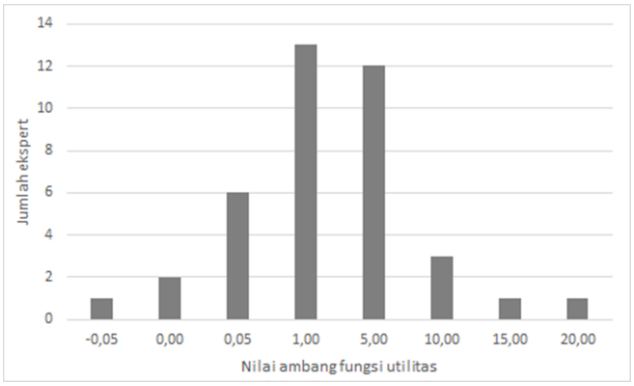
Validasi manual dan Gbr. 6 menunjukkan bahwa $thres = 5{,}0$ ditentukan sebagai nilai ambang, sehingga pakar domain IE cenderung lebih fokus ke dua topik riset dan pakar domain NLP ke 3-4 topik sebagai pilihan utama (Gbr. 5).

E. Evaluasi Hasil DCM

Regresi Logistik Multinomial untuk evaluasi model *Discrete Choice* menggunakan *package* mlogit dalam Bahasa R [2]. Berbagai skenario fungsi utilitas di Tabel VI telah dievaluasi. Terlihat bahwa model m_8 dengan dua ciri F_2, F_6 lebih baik dibanding model lain karena memiliki nilai *log.likelihood* yang lebih kecil.



Gbr. 5 Perbandingan persentase pakar pada domain IE dan NLP berdasarkan jumlah topik.



Gbr. 6 Histogram penentuan nilai fungsi utilitas sebagai acuan *threshold*.

Pemakaian semua ciri sebagai atribut alternatif pilihan memberikan model lebih baik (terlihat di m_1) dibanding hanya satu ciri dengan kinerja dinamisnya, yaitu model $m_2 \ldots m_7$. Di antara nilai terkait jumlah publikasi, model dengan ciri F_2 memiliki nilai *log.likelihood* kecil, sedangkan ciri F_6 yang terkait jumlah sitasi juga memiliki nilai *log.likelihood* lebih kecil. Lalu, model m_8 menggabungkan F_2, F_6, sehingga suatu topik riset dikatakan menjadi fokus pilihan jika kebiasaan pakar melakukan publikasi yang konsisten di bidang itu, fitur F_2, dan karyanya di bidang itu banyak mendapatkan sitasi pada suatu periode, fitur F_6. Perilaku menghasilkan publikasi lebih dipilih dalam bentuk kumulatif tanpa bobot, sedangkan perilaku mendapatkan sitasi lebih dipilih bentuk kumulatif dengan bobot.

Tabel VI adalah model hasil DCM dari data 52 pakar domain NLP dan IE dengan 30 fitur produktivitas serta kinerja dinamis. Tabel VII menunjukkan rata-rata hasil uji coba prediksi topik peneliti dari model hasil tersebut menggunakan *dataset* yang sama. Enam topik dengan peluang alternatif tertinggi dari hasil model m_1 diangggap sebagai data *ground-truth*. Rata-rata prediksi topik untuk 52 pakar dari model m_3, m_7, m_8 dengan enam prediksi teratas terlihat pada Tabel VII.

[2] https://cran.r-project.org/web/packages/mlogit /

TABEL VII
PRESISI PREDIKSI TOPIK DENGAN MODEL HASIL DCM

Model m_1		**Model m_3**		**Model m_7**		**Model m_8**	
Tpc	$p(Tpc)$	Tpc	$p(Tpc)$	Tpc	$p(Tpc)$	Tpc	$p(Tpc)$
T04	0,068	T13	0,124	T08	0,084	T13	0,115
T13	0,068	T29	0,089	T19	0,073	T18	0,077
T02	0,056	T18	0,061	T06	0,072	T04	0,058
T18	0,049	T04	0,060	T04	0,065	T19	0,058
T19	0,049	T14	0,059	T13	0,059	T27	0,058
Precision-5		60%		60%		80%	

TABEL VIII
PRESISI PREDIKSI TOPIK SEBAGAI PROBLEM KLASIFIKASI

	Nilai Presisi dengan range 0,0-1,0	
Algoritme	**60 kelas (topik dan status fokus utama)**	**2 kelas (status fokus utama)**
kNN	0,067	0,869
Random Forest	0,080	0,886
Naïve Bayes	0,022	0,800
Regresi Logistik	0,068	0,887

Precision-5 dari topik yang dapat diprediksi oleh setiap model dihitung berdasarkan data *ground-truth* dengan 60% untuk m_3, m_7 serta 80% untuk m_8. Nilai presisi sebagai indikator ketepatan prediksi tersebut menunjukkan bahwa DCM dengan fitur F_2, F_6 memberikan hasil prediksi topik lebih baik.

Prediksi topik dilakukan menggunakan algoritme klasifikasi k-NN, *Random Forest*, *Naive Bayes*, dan Regresi Logistik pada kakas bantu Orange[3] sebagai perbandingan. Nilai kelas adalah kombinasi topik dan status topik tersebut sebagai fokus utama (1 atau 0) sehingga terdapat 60 kelas (30 topik x 2 status). Metode *sampling* adalah *leave one out* dari 1.050 data. Namun, jumlah data per kelas tidak imbang dan kecenderungan ada pada kelas dengan status 0 di sebagian besar topik. Tabel VIII menunjukkan hasil presisi pada algoritme pembanding di atas.

Jika problem prediksi topik sebagai fokus utama dianggap klasifikasi dua kelas memberikan hasil presisi yang sama dengan DCM. Namun, target dua kelas adalah penyederhanaan masalah, sedangkan kondisi nyata membutuhkan 60 kelas yang menghasilkan presisi jauh lebih rendah. Beberapa proses telah dilakukan untuk validasi model prediksi, mulai dari penentuan topik, penentuan nilai ambang fungsi pilihan setelah pakar dipetakan ke topik, sampai penentuan kombinasi ciri yang tepat untuk pendekatan DCM. Terlihat bahwa problem penentuan fokus riset utama bagi pakar dengan variasi jumlah topik yang banyak dapat diselesaikan menggunakan DCM.

IV. KESIMPULAN

Berdasarkan hasil yang diperoleh, model *Discrete Choice* sebagai pendekatan model estimasi pada data bibliografi mendukung bahwa perilaku pakar memengaruhi penentuan fokus risetnya. Teori perilaku bahwa pakar aktif dalam melakukan publikasi sehingga hasil karyanya banyak disitasi dapat dibuktikan dengan DCM. Kedua perilaku tersebut memberikan ketepatan prediksi topik 80% lebih baik dibanding hanya berdasarkan salah satu perilaku yang memberikan prediksi 60%. Pada suatu periode, jika seorang pakar aktif melakukan publikasi di suatu topik riset dan karyanya di bidang itu banyak mendapat sitasi, maka pakar itu fokus ke topik tersebut. Persiapan data masukan untuk sesuai dengan karakter data model *Discrete Choice* memerlukan analisis tersendiri. Ekstraksi ciri hanya dilakukan dari sisi produktivitas pakar, sehingga informasi relasi pakar masih dapat dieksplorasi lebih lanjut dan diujicobakan pada domain penelitian yang tidak beririsan. Hal tersebut diperlukan untuk melihat kemungkinan domain topik sebagai pembeda dalam evaluasi model *Discrete Choice*.

[3] https://orange.biolab.si/

UCAPAN TERIMA KASIH

Terima kasih disampaikan kepada Lembaga Pengelola Dana Pendidikan (LPDP) atas dukungan dana yang telah diberikan dalam penelitian ini melalui skema Indonesian Education Scholarship for Doctoral Programs dengan nomor kontrak PRJ-4228/LPDP.3/2016 tahun fiskal 2017-2020.

Terima kasih juga disampaikan kepada Adhi Nurilham dan Amelia Sahira Rahma dari Laboratorium Komputasi Cerdas dan Visualisasi, Departemen Informatika, Institut Teknologi Sepuluh Nopember untuk persiapan dan validasi data.

REFERENSI

[1] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, dan L. Si, "Expertise Retrieval," *Found. Trends Inf. Retr.*, Vol. 6, No. 2–3, hal. 127–256, Feb. 2012.

[2] F. Xia, Z. Chen, W. Wang, J. Li, dan L.T. Yang, "MVCWalker: Random Walk-Based Most Valuable Collaborators Recommendation Exploiting Academic Factors," *IEEE Trans. Emerg. Top. Comput.*, Vol. 2, No. 3, hal. 364–375, Sep. 2014.

[3] F. Alarfaj, U. Kruschwitz, D. Hunter, and C. Fox, "Finding the Right Supervisor: Expert-Finding in a University Domain," *Proc. 2012 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Student Res. Work*, 2012., hal. 1–6.

[4] F. Xia, W. Wang, T.M. Bekele, dan H. Liu, "Big Scholarly Data: A Survey," *IEEE Trans. Big Data*, Vol. 3, No. 1, hal. 18–35, Mar. 2017.

[5] S. Lin, W. Hong, D. Wang, dan T. Li, "A Survey on Expert Finding Techniques," *J. Intell. Inf. Syst.*, Vol. 49, No. 2, hal. 255–279, Oct. 2017.

[6] H. Deng, I. King, dan M.R. Lyu, "Formal Models for Expert Finding on DBLP Bibliography Data," *2008 Eighth IEEE International Conference on Data Mining*, 2008, hal. 163–172.

[7] Z. Yang, J. Tang, B. Wang, J. Guo, dan J. Li, "Expert2Bólè : From Expert Finding to Bólè Search," *Proc. 15th ACM Conf. Knowl. Discov. data Min.*, 2009, hal. 1–4.

[8] J. Tang, "AMiner: Toward Understanding Big Scholar Data," *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, hal. 467.

[9] D. Purwitasari, C. Fatichah, I. K. E. Purnama, S. Sumpeno, dan M.H. Purnomo, "Inter-Departmental Research Collaboration Recommender System Based on Content Filtering in a Cold Start Problem," *2017 IEEE 10th International Workshop on Computational Intelligence and Applications, IWCIA 2017 - Proceedings*, 2017, hal. 177-184.

[10] L. Guo, X. Cai, F. Hao, D. Mu, C. Fang, dan L. Yang, "Exploiting Fine-Grained Co-Authorship for Personalized Citation

Recommendation," *IEEE Access*, Vol. 5, hal. 12714–12725, 2017.

[11] G. Panagopoulos, G. Tsatsaronis, dan I. Varlamis, "Detecting Rising Stars in Dynamic Collaborative Networks," *J. Informetr.*, Vol. 11, No. 1, hal. 198–222, 2017.

[12] H. Jiang, "A Nested Logit-Based Approach to Measuring Air Shopping Screen Quality and Predicting Market Share," *J. Revenue Pricing Manag.*, Vol. 8, No. 2, hal. 134–147, Mar. 2009.

[13] H. Jiang, X. Qi, dan H. Sun, "Choice-Based Recommender Systems: A Unified Approach to Achieving Relevancy and Diversity," *Oper. Res.*, Vol. 62, No. 5, hal. 973–993, Oct. 2014.

[14] M. Paredes, E. Hemberg, U. O'Reilly, dan C. Zegras, "Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction?," *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 2017, hal. 780–785.

[15] A. Nurilham, D. Purwitasari, dan C. Fatichah, "Ekstraksi Frasa pada Pelabelan Kelompok Artikel Ilmiah dengan Penggabungan Klaster berdasarkan MaximumCommonSubgraph," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 7, No. 3, hal. 249-257, 2018.

[16] M. Ben-Akiva dan S.R. Lerman, *Discrete Choice Analysis : Theory and Application to Travel Demand*, Cambridge, USA: MIT Press, 1985.

[17] V. Aguirregabiria dan P. Mira, "Dynamic Discrete Choice Structural Models: A Survey," *J. Econom.*, Vol. 156, No. 1, hal. 38–67, 2010.

[18] E. Lancsar, J. Louviere, C. Donaldson, G. Currie, dan L. Burgess, "Best Worst Discrete Choice Experiments in Health: Methods and an Application," *Soc. Sci. Med.*, Vol. 76, hal. 74–82, 2013.

[19] G. Antonini, M. Bierlaire, dan M. Weber, "Discrete Choice Models of Pedestrian Walking Behavior," *Transp. Res. Part B Methodol.*, Vol. 40, No. 8, hal. 667–687, 2006.

[20] D. McFadden, "The Measurement of Urban Travel Demand," *J. Public Econ.*, Vol. 3, No. 4, hal. 303–328, 1974.

[21] D.R. Radev, P. Muthukrishnan, V. Qazvinian, dan A. Abu-Jbara, "The ACL Anthology Network Corpus," *Lang. Resour. Eval.*, Vol. 47, No. 4, hal. 919–944, Dec. 2013.

[22] J. Santoso, A.D.B. Soetiono, Gunawan, E. Setyati, E.M. Yuniarno, M. Hariadi, dan M.H. Purnomo, "Self-Training Naive Bayes Berbasis Word2Vec untuk Kategorisasi Berita Bahasa Indonesia," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 7, No. 2, hal. 158–166, 2018.

[23] A. Zaini, M. A. Muslim, dan Wijono, "Pengelompokan Artikel Berbahasa Indonesia Berdasarkan Struktur Laten Menggunakan Pendekatan Self Organizing Map," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 6, No. 3, hal. 259–267, 2017.

[24] O. Somantri dan M. Khambali, "Feature Selection Klasifikasi Kategori Cerita Pendek Menggunakan Naïve Bayes dan Algoritme Genetika," *J. Nas. Tek. Elektro dan Teknol. Inf.*, Vol. 6, No. 3, hal. 301–306, 2017.

[25] R. Řehůřek dan P. Sojka, "Software Framework for Topic Modelling with Large Corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, hal. 45–50.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, dan D. Cournapeau, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, Vol. 12, hal. 2825–2830, 2011.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, Vol. 20, hal. 53–65, 1987.