Cross-Domain Topic Learning Berbasis Frase untuk Pemodelan Topik pada Rekomendasi Kolaborasi Penelitian

Vit Zuraida[1], Diana Purwitasari[2], Chastine Fatichah[3]
[1,2,3]Departemen Informatika, Fakultas Teknologi Informasi dan Komunikasi,
Institut Teknologi Sepuluh Nopember
[1]vit.zuraida@gmail.com

Abstract. *Cross-domain collaboration recommendation can be inferred from texts of published research documents such as titles, abstracts, and bibliographies. Extracting researchers' topics is necessary for topic modeling in the recommendation system. In bag-of-words approach for topic modeling, word sequence is often disregarded. Therefore, phrases are preferable than single words for representing topics. We propose cross-domain collaboration recommendation system using phrase-based cross domain topic learning. The proposed method considers the rarity of relevant cross-domain collaboration topics so that it does not adversely affect the accuracy of recommendation results. The proposed method consists of three main steps: (1) transforming documents from bag-of-words to bag-of-phrases representation, (2) topic modeling to learn the probability of researchers' topic interests, and (3) recommendation ranking using random walk with restart. Experiments on Visualization and Data Mining domain from AMiner dataset shows that phrase-based CTL performs better than CTL based on bag-of-words. There is ± 10% improvement of precision value in the top 10 recommendations and ± 5% improvement in the top 20 recommendations.*

Keywords: *Cross-Domain Collaboration Recommendation, Topic Model, Random Walk*

Abstrak. *Rekomendasi kolaborasi penelitian antardomain dapat diperoleh melalui dokumen publikasi ilmiah seperti judul, abstrak, dan bibliografi. Oleh karena itu, proses ekstraksi topik riset dari seorang peneliti merupakan tahapan penting. Model topik berbasis kata belum dapat merepresentasikan topik dengan baik sebab urutan kata pada dokumen tidak diperhitungkan. Penelitian ini mengusulkan sistem rekomendasi kolaborasi antardomain dengan metode Cross-Domain Topic Learning (CTL) Berbasis Frase. CTL Berbasis Frase terdiri dari tiga fase utama: (1) transformasi dokumen dari format bag-of-words menjadi bag-of-phrases, (2) pemodelan topik terhadap frase yang sudah dibentuk untuk mengetahui distribusi probabilitas keterkaitan peneliti dengan topik, (3) perangkingan rekomendasi kolaborasi dengan random walk with restart. Pengujian sistem terhadap domain Visualization dan Data Mining pada dataset AMiner menunjukkan bahwa CTL Berbasis Frase lebih baik daripada CTL berbasis kata. Terdapat pengingkatan nilai precision sebesar ±10% pada 10 rekomendasi teratas dan ±5% pada 20 rekomendasi teratas.*

Kata Kunci: *Rekomendasi Kolaborasi Antardomain, Model Topik, Random Walk*

1. Pendahuluan

Penelitian merupakan salah satu aspek penting dalam pengembangan bidang keilmuan (Pangestu, 2017). Berbagai penelitian yang telah dilakukan tidak hanya terkonsentrasi pada suatu domain tertentu, namun bisa jadi merupakan kolaborasi dari beberapa domain. Kolaborasi antardomain menggabungkan beragam keahlian dan terbukti efektif dalam pemecahan masalah kompleks yang memunculkan hasil inovatif baik secara teoritis maupun aplikatif (Liang, 2017). Sebagai contoh kolaborasi penelitian dalam efisiensi energi (Hemptinne, 2017), teknik biomedis (Khamidah, 2018)(Santoso, 2017), dan konservasi alam (Mitchell, 2017). Rekomendasi kolaborasi bertujuan membantu peneliti dalam menemukan kolaborator yang sesuai pada domain sama, antardomain, antardepartemen dalam institusi (Purwitasari, 2017), maupun institusi dengan industri (Wang, 2017). Metode sistem rekomendasi terbagi menjadi kategori *collaborative filtering* yang dibangun berdasarkan persamaan pengguna yaitu peneliti, dan *content-based* berdasarkan persamaan *item* yaitu isi publikasi penelitian (Kang, 2015).

Beberapa metode telah dikembangkan dalam identifikasi dan rekomendasi penelitian antardomain, seperti HyClass yang menggunakan similaritas semantik pada kata dalam publikasi dengan memanfaatkan taksonomi MeSH (Kang, 2015). Selain itu Osuna (2017) membentuk rekomendasi kolaborasi dengan klasifikasi peneliti berdasarkan *research footprint* yaitu kata kunci, konsep, atau judul dan abstrak. Liang (2017) memberikan rekomendasi melalui metode *clustering* dengan menghitung nilai persamaan antartopik.

Sistem rekomendasi kolaborasi antardomain memberikan rekomendasi kolaborator dari domain tertentu untuk seorang peneliti dari domain lain. Model rekomendasi dibentuk berdasarkan data judul dan abstrak dari publikasi yang dimiliki oleh peneliti tersebut. Sebagai contoh peneliti dari domain *data mining*, disebut domain asal, ingin menemukan kolaborator pada domain *visualization*, disebut domain target. Artikel ilmiah yang dipublikasikan pada domain *data mining* disebut publikasi domain asal dan artikel pada domain *visualization* disebut publikasi domain target. Seorang peneliti dianggap sebagai peneliti suatu domain tertentu jika memiliki publikasi pada domain tersebut.

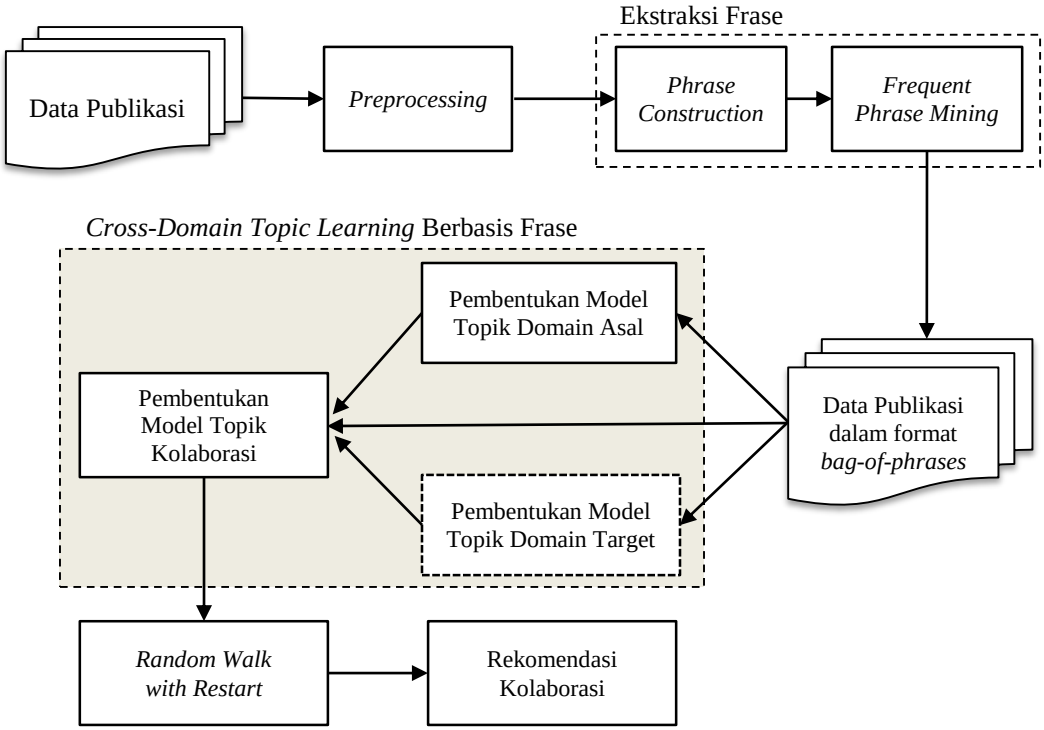
Koneksi tersebar karena relasi antar peneliti pada domain berbeda membuat proses menemukan kolaborator yang sesuai menjadi tidak mudah (Tang, 2012). Hal tersebut dipengaruhi oleh perbedaan keahlian yang menyebabkan perbedaan terminologi pada domain. Selain itu relevansi topik untuk kolaborasi antardomain menunjukkan hanya 9% dari seluruh kemungkinan pasangan domain yang diuji memiliki kolaborasi penelitian. Ketiga permasalahan tersebut mendorong Tang mengusulkan metode *Cross-Domain Topic Learning* (CTL) yang mempertimbangkan adanya topik eksklusif terkait dengan satu domain sehingga tidak relevan untuk kolaborasi antardomain. CTL memastikan kelangkaan topik relevan tidak berpengaruh pada akurasi sistem rekomendasi. CTL adalah model rekomendasi kolaborasi yang dibentuk berdasarkan persamaan konten (judul dan abstrak dari dokumen penelitian). Oleh karena itu, proses ekstraksi topik riset dari masing-masing peneliti merupakan tahapan yang penting. Banyak pemodelan topik untuk rekomendasi kolaborasi yang menggunakan *Latent Dirichlet Allocation* (LDA) dengan pendekatan kata tunggal (*bag-of-words*) (Liang, 2017). Pemodelan topik dengan frasa dapat menghasilkan perbedaan seperti kata "*text*" dalam frase '*text preprocessing*" dan "*text visualization*". Model topik dengan *bag-of-words* mengasumsikan bahwa kata berdiri secara independen. Makna frase tidak selalu dapat disimpulkan dari makna kata penyusunnya (Schone, 2001). Frase dengan kombinasi kata tertentu menyimpan informasi lebih penting dibanding dengan gabungan nilai informasi dari setiap kata penyusun (Wang, 2007). Model topik berbasis frase dikembangkan karena lebih representatif dalam menentukan topik seperti frase "*automatic keyword extraction*" merepresentasikan topik *text mining* lebih baik dibanding kata "*automatic*", "*keyword*", dan "*extraction*" secara terpisah. Beberapa metode model topik berbasis frase antara lain KERT (Danilevsky, 2013), TurboTopic (Blei, 2009), *Phrase-Discovering* LDA (Lindsey, 2012), dan ToPMine (El-Kishky, 2014).

Penelitian ini mengusulkan rekomendasi kolaborasi penelitian antardomain dengan model topik berbasis frase yang dikembangkan dari *Cross-Domain Topic Learning*. Transformasi dokumen dalam format *bag-of-words* menjadi *bag-of-phrases* pada pemodelan topik dilakukan dengan metode ToPMine. Kemudian *bag-of-phrases* yang terbentuk menjadi masukan dalam pemodelan rekomendasi kolaborasi antardomain sebagai kontribusi pada makalah ini.

Bahasan selanjutnya terbagi dalam beberapa bagian. Bagian Metode Usulan akan memaparkan tahapan transformasi dokumen publikasi menjadi *bag-of-words* dan proses pemodelan topik hingga diperoleh rekomendasi kolaborator. Bagian Hasil dan Pengujian menjelaskan mengenai persiapan data serta skenarion dan hasil uji coba. Sebagai penutup, bagian Kesimpulan juga menjelaskan mengenai penelitian lanjutan yang akan dilakukan.

2. Metode Usulan

Usulan model sistem rekomendasi kolaborasi penelitian antardomain terdiri dari empat fase utama pada Gambar 1. Fase pertama adalah *preprocessing* terhadap judul dan abstrak dari publikasi. Selanjutnya dilakukan fase ekstraksi frase untuk melakukan transformasi publikasi dari *bag-of-words* menjadi *bag-of-phrases*. Berikutnya pada fase *Cross-Domain Topic Learning* Berbasis Frase dilakukan ekstraksi topik menggunakan frase yang telah dibentuk. Pada fase akhir, *Random Walk with Restart* digunakan untuk memperoleh rekomendasi kolaborator yang sesuai untuk seorang peneliti.



Gambar 1. Desain Sistem Rekomendasi Kolaborasi Penelitian Antardomain

2.1. Fase *Preprocessing*

Fase *preprocessing* meliputi empat tahapan yaitu (1) *case folding* yang mengubah alfabet pada teks input menjadi huruf kecil, (2) *tokenization* untuk tokenisasi dokumen menjadi kata dengan pemisah spasi maupun tanda baca, (3) *stopword removal* untuk eliminasi kata yang memiliki frekuensi kemunculan tinggi pada dokumen seperti "*the*", "*and*", "*to*", dan (4) *stemming* untuk konversi token kata berimbuhan menjadi kata dasar. *Stemmer* yang digunakan dalam penelitian ini adalah *Porter Stemmer* karena dokumen publikasi tertulis dalam Bahasa Inggris. Masukan pada fase ini adalah koleksi dokumen beserta keterangan peneliti yang menjadi penulisnya. Luaran fase *preprocessing* adalah kumpulan token dalam bentuk kata dasar untuk setiap dokumen.

2.2. Fase Ekstraksi Frase

Fase esktraksi frase bertujuan untuk transformasi dokumen dari *bag-of-words* menjadi *bag-of-phrases* yang terdiri dari dua tahap, yaitu *frequent phrase mining* dan *phrase construction.* Tahap *frequent phrase mining* memperoleh frase-frase dengan jumlah kemunculan pada koleksi dokumen lebih tinggi daripada batas minimal kemuncul yang ditentukan. Nilai batas tersebut didefinisikan sebagai *minimum support*. Tahapan dalam *frequent phrase mining*:

(a) Ekstraksi frase dari suatu dokumen dimulai dari kata pertama pada dokumen tersebut. Posisi kata pada dokumen setelah tahap *preprocessing* itu ditandai dengan nilai indeks aktif. Frase dapat terdiri dari *n* kata. Dilakukan penyimpanan nilai indeks aktif untuk frase dengan panjang *n*. Sebagai inisialisasi, pada iterasi pertama panjang frase *n* diset 1.

(b) Penghitungan frekuensi kemunculan setiap frase dengan panjang *n*. Eliminasi frase dengan frekuensi kemunculan yang tidak memenuhi *minimum support* dengan cara menghapus indeks frase tersebut dari himpunan indeks aktif agar tidak diperhitungkan pada iterasi selanjutnya.

(c) Eliminasi dokumen publikasi yang tidak memiliki *frequent phrase* dengan panjang n. Jika suatu dokumen tidak memiliki *frequent phrase* dengan panjang *n*, maka dokumen tersebut dipastikan tidak memiliki *frequent phrase* dengan panjang lebih besar dari *n*.

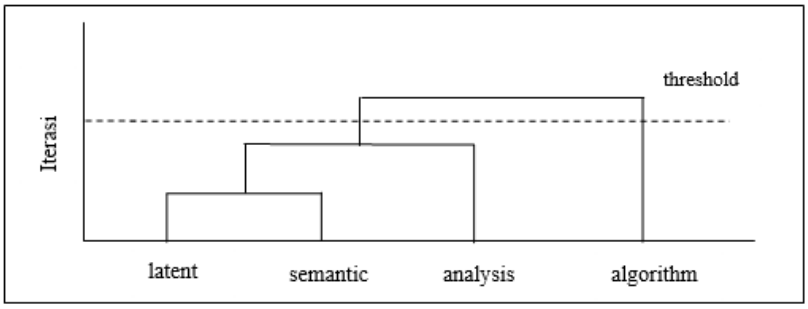
Ketiga langkah tersebut terus dilakukan hingga tidak ada lagi dokumen yang bisa diproses. Hasil tahapan *frequent phrase mining* adalah *frequent phrase* beserta jumlah kemunculannya.

Tahap *phrase construction* mengeliminasi frase-frase yang diidentifikasi sebagai *frequent phrase* karena memenuhi *minimum support* namun sebenarnya bukan termasuk frase yang valid. Langkah-langkah pada *phrase construction* terdiri dari tiga tahapan sebagai berikut:
(a) Penghitungan nilai *significance score* untuk setiap pasang frase P_1 dan P_2 yang berurutan (Lihat Persamaan (1)). Notasi $f(P_1 \oplus P_2)$ menyatakan frekuensi kemunculan rangkaian frase yang dibentuk P_1 dan P_2 sedangkan notasi $\mu_0\ (P_1, P_2)$ adalah rata-rata frekuensi kemunculan rangkaian P_1 dan P_2 untuk mengenali bahwa frase P_1 dan P_2 adalah frase valid. Nilai $\mu_0\ (P_1, P_2)$ dihitung dengan Persamaan (2) dengan L adalah jumlah token pada korpus dan $p(P) = \frac{f(P)}{L}$ adalah estimasi nilai probabilitas kemunculan frase dalam korpus.
(b) Penentuan nilai pasangan frase dengan nilai *significance score* terbaik dari seluruh pasangan frase. Jika nilai *significance score* terbaik lebih besar dari *threshold*, maka pasangan frase tersebut digabungkan menjadi frase baru. Sebaliknya, jika nilai *significance score* lebih kecil daripada *threshold*, maka proses *phrase construction* dihentikan.
(c) Penghitungan nilai *significance score* untuk frase baru P_1 dan P_2. Ketiga langkah ini terus dilakukan hingga nilai *significance score* terbaik tidak memenuhi *threshold* atau saat semua frase sudah digabungkan. Contoh proses *phrase construction* ditunjukkan dalam Gambar 2.

Berdasarkan hasil proses *frequent phrase mining*, frase "*latent semantic", "latent semantic analysis"* dan *"latent semantic analysis algorithm"* termasuk *frequent phrase* karena kemunculannya memenuhi *minimum support.* Namun pada tahap *phrase construction,* frase valid yang diperoleh berdasarkan nilai *significance score* adalah *"latent semantic analysis".* Hasil akhir dari tahap ini adalah dokumen publikasi dalam format *bag-of-words.*

$$sig(P_1, P_2) \approx \frac{f(P_1 \oplus P_2) - \mu_0\ (P_1, P_2)}{\sqrt{f(P_1 \oplus P_2)}} \quad (1)$$

$$\mu_0\ (P_1, P_2) = L\ x\ p(P_1)\ x\ p(P_2) \quad (2)$$



Gambar 2. Contoh Proses *Phrase Construction*

.

2.3. Fase *Cross-Domain Topic Learning* Berbasis Frase

Hasil transformasi *bag-of-words* pada judul dan abstrak publikasi menjadi objek dalam pembentukan model topik untuk setiap peneliti dan pasangan peneliti dari domain asal dan domain target. Masukan pada fase ini adalah judul dan abstrak dari publikasi ilmiah yang sudah ditransformasikan ke dalam format *bag-of-phrases*. Pada fase ini, publikasi dibagi menjadi tiga kategori, yaitu publikasi dari domain asal, publikasi dari domain target, dan publikasi yang merupakan kolaborasi dari kedua domain tersebut.
(a) Untuk tahap awal, dilakukan pembentukan model topik pada domain asal dan target, sesuai dengan distribusi multinomial $p(\theta_v|\alpha)$ dan $p(\theta'_{v'}|\alpha)$. Definisi notasi ditunjukkan pada Tabel 1.
(b) Tahap selanjutnya adalah pemodelan publikasi yang termasuk kategori kolaborasi antardomain. Untuk setiap C_{dg}, nilai *s* diperoleh berdasarkan *beta distribution* $p(s|d) \sim beta(\gamma, \gamma_t)$. Jika *s* bernilai 1 yang berarti publikasi pada satu domain, maka penulis *v* (atau *v'*) dipilih berdasarkan *uniform distribution.* Selanjutnya *sampling* dilakukan terhadap C_{dg} dengan topik z_{dg} spesifik terhadap user *v* sesuai dengan θ_v. Jika *s* bernilai 0, dipilih pasangan kolaborasi penulis (*v, v'*) dan distribusi

multinomial $\vartheta_{vv'}$ dibentuk dengan menggabungkan θ_v dan $\theta_{v'}$. Penggabungan kedua model topik dilakukan dengan terlebih dahulu menyamakan dimensi keduanya. Selanjutnya adalah *sampling* C_{dg} dari topik kolaborasi z_{dg} berdasarkan distribusi $\vartheta_{vv'}$ yang baru. Berbeda dengan LDA terhadap *bag-of-words*, perhitungan nilai probabilitas posterior pada CTL berbasis frase memiliki batasan bahwa setiap kata pada frase yang sama akan dikaitkan pada topik yang sama. Oleh karena itu, proses pengambilan sampel nilai *s* dan *z* dilakukan sekali untuk setiap frase. Notasi yang digunakan dalam fase ini dirangkum dalam Tabel 1.

Tabel 1. Notasi CTL Berbasis Frase

Simbol	Deskripsi
T	Himpunan topik
d	Dokumen kolaborasi
A_d	Himpunan penulis untuk dokumen *d*
X_{dg}	Frase ke *g* pada dokumen *d*
x_{dgj}	Token ke-*j* pada dokumen *d* frase ke *g*
z_{dgj}	Topik yang dikaitkan dengan x_{dgj}
$C_{d,g}$	$\{z_{d,g,j}\}_{j=1}^{\lvert X_{d,g}\rvert}$ yaitu kumpulan topik pada frase ke *g* di dokumen *d*
s_{dgj}	Bernilai 1 jika x_{dgj} adalah kata yang termasuk *single domain* atau dan 0 jika termasuk *cross domain*
θ dan θ'	Distribusi multinomial dari topik pada domain asal dan domain target
θ_v	Distribusi multinomial dari topik spesifik terhadap penulis *v*
$\vartheta_{vv'}$	Distribusi multinomial dari topik spesifik terhadap pasangan penulis (*v, v'*)
ϕ_z	Distribusi multinomial dari kata spesifik terhadap topik *z*
α, β	Parameter Dirichlet
λ	Parameter untuk *sampling* variabel *s*
γ, γ_t	Parameter beta untuk menghasilkan nilai λ

Metode ini menggunakan *Gibbs Sampling* untuk mengestimasi nilai parameter $\{\theta, \theta', \vartheta, \phi, \lambda\}$. Probabilitas posterior pada *z* (atau *z'*) untuk setiap kata pada dokumen publikasi oleh penulis dari *single* domain menggunakan persamaan (3). Hasil perhitungan ini akan mempengaruhi nilai θ (atau θ').

$$P(C_{dg} = z|x,.) = \sum_{j=1}^{X_{dg}} (\alpha + n_{vz_{dgj}}^{-dgj} + j - 1) \times \frac{m_{z_{dgj}x_{dgj}}^{-dgj} + \beta}{\sum_x \left(m_{z_{dgj}x}^{-dgj} + \beta\right) + j - 1} \quad (3)$$

n_{vz} : Jumlah berapa kali topik *z* menjadi label untuk penulis *v*
m_{zx} : Jumlah berapa kali kata *x* dilabeli topik *z*
n^{-dgj} : Notasi *–dgj* berarti jumlah tidak memperhitungkan item yang saat ini diproses

Probabilitas posterior pada *s* dihitung dengan persamaan (4) dan hasilnya digunakan untuk memperoleh nilai parameter $\theta, \theta', \vartheta$ dengan persamaan (5). Selanjutnya nilai ϕ, λ dapat disimpulkan dari model topik yang dibentuk. Persamaan (4) bisa disesuaikan untuk $P(s_{dg} = 1|.)$. Perubahan yang perlu diperhatikan adalah mengganti $\left(n_{vz_{dgj}} + n_{v'z_{dgj}}\right)$ dengan penulis tunggal $n_{vz_{dgj}}$ atau $n_{v'z_{dgj}}$.

$$P(s_{dg} = 0|z,.) = \sum_{j=1}^{X_{dg}} \left(\alpha + n_{vv'z_{dgj}}^{-dgj} + \left(n_{vz_{dgj}} + n_{v'z_{dgj}}\right) + j - 1\right) \times \frac{n_{ds_0}^{-dgj} + \gamma_t}{n_{ds_0}^{-dgj} + n_{ds_1}^{-dgj} + \gamma_t + \gamma} \quad (4)$$

n_{ds_0} : Jumlah berapa kali 0 menjadi sampel pada dokumen *d*
(v, v') : Pasangan penulis yang dipilih untuk suatu x_{dgj}
$n_{vv'z}$: Jumlah berapa kali topik *z* dijadikan label untuk (v, v')

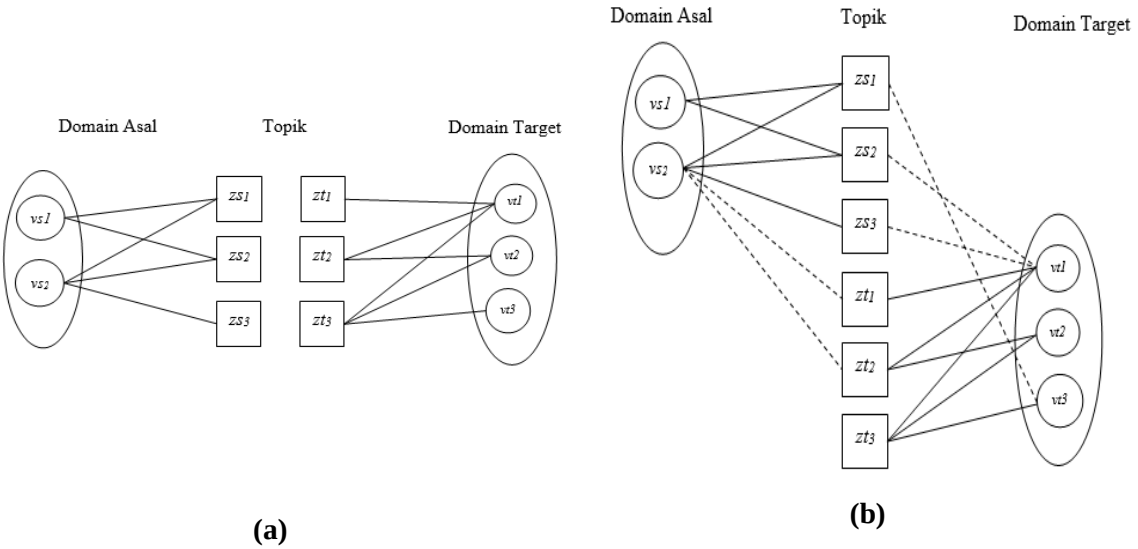
ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

Probabilitas posterior topik *z* didefinisikan pada persamaan (5) sebagai berikut:

$$P(C_{dg} = z|s_{dg} = 0, x, .) = \sum_{j=1}^{X_{dg}} (\alpha + n_{vv'z_{dgj}}^{-dgj} + (n_{vz_{dgj}} + n_{v'z_{dgj}})) \times \frac{m_{z_{dgj}x_{dgj}}^{-dgj} + m_{z_{dgj}x_{dgj}} + m'_{z_{dgj}x_{dgj}} + \beta}{\sum_x \left(m_{z_{dgj}x}^{-dgj} + m_{z_{dgj}x} + m'_{z_{dgj}x} + \beta\right)} \quad (5)$$

m_{zx}^{-dgj}: Jumlah berapa kali kata *x* dilabeli topik *z* pada publikasi kolaborasi
m_{zx} : Jumlah berapa kali kata *x* dilabeli topik *z* di publikasi domain asal
m'_{zx} : Jumlah berapa kali kata *x* dilabeli topik *z* di publikasi domain target

Pada Gambar 4(a) ditunjukkan gambaran hasil pemodelan topik pada masing-masing domain asal dan target. Pada tahap ini, peneliti pada domain asal hanya memiliki distribusi probabilitas terhadap topik domain asal. Setelah dilakukan *Cross-Domain Topic Learning* terhadap publikasi kolaborasi, peneliti pada domain asal juga akan memiliki distribusi probabilitas terhadap topik pada domain target, dan begitu juga sebaliknya, seperti yang ditunjukkan pada Gambar 4(b).



Gambar 3. Pemodelan Topik (a) Pemodelan pada Publikasi Domain Asal dan Target (b) Pemodelan pada Publikasi Kolaborasi

2.4. Fase Perankingan dengan *Random Walk with Restart*

Luaran dari fase *Cross-Domain Topic Learning* Berbasis Frase berupa distribusi probabilitas topik untuk setiap peneliti selanjutnya dijadikan dasar dalam perangkingan rekomendasi kolaborasi. Perankingan dilakukan dalam beberapa tahapan sebagai berikut: (1) Pembentukan *graph* berdasarkan hasil *CTL*. Suatu *node* peneliti dihubungkan dengan *node* topik tertentu jika memiliki probabilitas posterior $P(z|s = 0, .)$ lebih besar dari *threshold.* Semakin kecil *threshold* yang diberlakukan maka *graph* yang terbentuk akan semakin padat. (2) Penghitungan nilai keterkaitan antara peneliti-peneliti pada domain target dengan *query node* peneliti pada domain asal. Nilai keterkaitan diperoleh berdasarkan algoritma *Random Walk with Restart* hingga dicapai konvergensi. (3) Perangkingan penulis pada domain target berdasarkan nilai keterkaitan terbesar untuk menentukan rekomendasi kolaborasi yang diberikan.

3. Eksperimen

Uji coba sistem dilakukan untuk mengukur *precision* dan *recall* dari model rekomendasi kolaborasi antardomain terhadap data uji, yaitu judul, abstrak, dan peneliti dari publikasi pada sepasang domain yang dipilih dari dataset *Arnetminer.org*, yaitu *Visualization* sebagai domain asal dan *Data Mining* sebagai domain target. Domain asal merupakan domain peneliti yang dijadikan sebagai *input query* sedangkan domain target adalah domain dari kolaborator yang hendak dicari.

Fase uji coba dilakukan dengan membagi dataset menjadi dua bagian. Penelitian yang dipublikasikan sebelum tahun 2001 dijadikan data latih dan penelitian yang dipublikasikan pada tahun 2001 dan berikutnya dijadikan data uji. Jika rekomendasi yang diberikan sistem berdasarkan data latih kemudian terlaksana, maka rekomendasi ini dianggap rekomendasi yang benar, dan begitu juga sebaliknya. Tabel 2 menunjukkan informasi mengenai dataset yang dipakai.

Tabel 2. Informasi Data Uji Coba

Keterangan	Jumlah
Jumlah peneliti domain asal	5399
Jumlah peneliti domain target	3274
Jumlah publikasi domain asal	3862
Jumlah publikasi domain target	2190
Jumlah publikasi kolaborasi	535
Jumlah publikasi kolaborasi data latih	210
Jumlah publikasi kolaborasi data uji	325
Jumlah peneliti domain asal yang memiliki kolaborasi penelitian antardomain pada data latih	256
Jumlah peneliti domain asal yang memiliki minimal 3 kolaborasi penelitian antardomain pada data uji	99
Jumlah peneliti domain asal yang memiliki kolaborasi penelitian antardomain pada data latih dan minimal 3 kolaborasi penelitian pada data uji	57

3.1. Pengujian terhadap *Minimum Support*

Uji coba ini dilakukan untuk mengetahui pengaruh nilai *minimum support* pada proses ekstraksi frase terhadap hasil rekomendasi. Uji coba dilakukan dengan nilai *minimum support* 30, 40, 50, dan 60. Hasil uji coba ditunjukkan pada Tabel 3. Nilai *precision* dan *recall* tidak mengikuti perubahan nilai *minimum support* dengan pola tertentu. Nilai *precision* dan *recall* terbaik dihasilkan dengan *minimum support* 50 sedangkan nilai *precision* dan *recall* terendah dihasilkan dengan *minimum support* 60.

Tabel 3. Hasil Pengujian terhadap *Minimum Support*

Minimum Support	Top 10		Top 20		Top 100	
	Precision Terbaik (%)	P@10 (%)	Precision Terbaik (%)	P@20 (%)	Recall Terbaik (%)	R@100 (%)
30	60.00	9.12	40.00	6.49	100.00	37.38
40	50.00	7.72	30.00	6.23	100.00	36.42
50	**60.00**	**9.47**	**40.00**	**6.67**	**100.00**	**41.53**
60	30.00	7.72	40.00	5.96	100.00	33.32

3.2. Pengujian terhadap Jumlah Topik

Pengujian ini bertujuan untuk mengetahui jumlah topik yang menghasilkan rekomendasi terbaik. Pengujian dilakukan dengan jumlah topik 30, 35, 40, 45, dan 50 dengan beberapa variasi nilai *minimum support.* Hasil uji coba ditunjukkan pada Tabel 4. Sama halnya dengan pengujian pada *minimum support,* hasil pengujian terhadap jumlah topik juga tidak spesifik mengikuti tren tertentu, namun *precision* dan *recall* yang terbaik umumnya dicapai pada jumlah topik yang lebih besar.

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

Tabel 4. Hasil Pengujian terhadap Jumlah Topik

Min Support	Jumlah Topik	Top 10		Top 20		Top 100	
		Precision Terbaik (%)	P@10 (%)	Precision Terbaik (%)	P@20 (%)	Recall Terbaik (%)	R@100 (%)
30	30	40.00	6.84	35.00	**6.67**	100.00	30.10
	35	50.00	8.77	50.00	5.96	100.00	31.19
	40	60.00	**9.47**	40.00	6.40	100.00	34.14
	45	**70.00**	9.12	45.00	6.23	100.00	**34.92**
	50	60.00	9.12	40.00	6.49	100.00	37.38
40	30	50.00	6.49	40.00	5.79	100.00	33.08
	35	50.00	8.42	35.00	5.44	100.00	34.11
	40	40.00	8.77	**45.00**	5.61	100.00	35.30
	45	**60.00**	**9.30**	40.00	6.05	100.00	33.97
	50	50.00	7.72	30.00	**6.23**	100.00	**36.42**
50	30	50.00	7.19	40.00	5.00	100.00	29.51
	35	**70.00**	9.30	50.00	5.79	100.00	33.50
	40	40.00	8.25	**45.00**	5.61	100.00	35.10
	45	30.00	9.12	35.00	6.23	100.00	35.19
	50	60.00	**9.47**	40.00	**6.67**	100.00	**41.53**

3.3. Perbandingan CTL dan CTL Berbasis Frase

Uji coba juga dilakukan untuk membandingkan performa CTL dengan CTL Berbasis Frase. Hasil uji coba pada Tabel 5 menunjukkan bahwa CTL Berbasis Frase menghasilkan nilai *precision* dan *recall* rata-rata lebih baik dibandingkan dengan CTL. Kinerja CTL lebih baik pada beberapa kasus seperti pada nilai *precision* pada 10 dan 20 rekomendasi terbaik dengan jumlah topik 50.

Tabel 5. Hasil Uji Coba Perbandingan CTL dan CTL Berbasis Frase

Jumlah Topik	Top 10				Top 20				Top 100			
	Precision Terbaik (%)		P@10 (%)		Precision Terbaik (%)		P@20 (%)		Recall Terbaik (%)		R@100 (%)	
	CTL	CTLBF	CTL	CTLBF	CTL	CTLBF	CTL	CTLBF	CTL	CTLBF	CTL	CTLBF
30	20.00	**50.00**	6.49	**7.19**	30.00	**40.00**	**5.09**	5.00	100	100	**32.95**	29.51
35	30.00	**70.00**	6.84	**9.30**	30.00	**50.00**	4.91	**5.79**	100	100	29.62	**33.50**
40	40.00	**60.00**	8.42	**9.47**	**40.00**	**40.00**	5.44	**6.40**	100	100	**35.96**	34.14
45	60.00	**60.00**	8.77	**9.30**	**45.00**	40.00	5.88	**6.05**	100	100	32.69	**33.97**
50	50.00	**60.00**	**10.35**	9.47	**45.00**	40.00	**7.11**	6.67	100	100	37.40	**41.53**
Rata-rata	40.00	**60.00**	8.17	**8.95**	38.00	**42.00**	5.69	**5.98**	100	100	33.72	**34.53**

Rata-rata *precision* dan *recall* pada hasil rekomendasi baik dengan metode CTL maupun CTL Berbasis Frase yang diberikan sangat rendah. Hal ini disebabkan karena tidak tersedianya cukup publikasi dari masing-masing peneliti pada data latih sehingga pembentukan distribusi probabilitas topik untuk peneliti tersebut kurang representatif. **Error! Reference source not found.** menunjukkan contoh daftar peneliti yang memiliki nilai *precision* 0% pada CTL berbasis frase dengan jumlah topik 50.

Tabel 6. Peneliti dengan Nilai *Precision* 0%

No	Nama	Jumlah Kolaborasi pada Data Uji	Jumlah Publikasi pada Data Latih
1.	Paul R. Cohen	3	1
2.	Steve Lawrence	6	1
3.	David Hart	4	1
4.	Bernhard Scholkopf	5	1

Perbedaan hasil rekomendasi CTL Berbasis Frase dan CTL umumnya ditunjukkan oleh peneliti dengan publikasi yang mengandung banyak *frequent phrase*, salah satunya ditunjukkan pada perbedaan nilai *precision* untuk peneliti "Vincent Y. Lum". Rekomendasi dengan CTL menghasilkan *precision* 0% sedangkan dengan CTL Berbasis Frase memperoleh *precision* 40%. Berdasakran analisa, salah satu publikasi "Vincent Y. Lum" pada data latih memiliki delapan *frequent phrase.* Dengan CTL Berbasis Frase, kata "*natural*" yang merupakan bagian dari frase "*natural language*" akan memiliki peluang lebih besar untuk dikaitkan dengan topik "*text processing*" daripada kata "*natural*" yang di-*sampling* per kata. Dengan model topik yang lebih representatif, rekomendasi kolaborator yang diberikan akan lebih baik pula.

4. Kesimpulan

Penelitian ini mengusulkan *Cross-Domain Topic Learning* Berbasis Frase sebagai metode pemodelan topik dalam sistem rekomendasi kolaborasi antardomain. Berdasarkan hasil pengujian, CTL Berbasis Frase menghasilkan nilai *precision* dan *recall* terbaik pada nilai *minimum support* 50 dan jumlah topik 50. Meski demikian, nilai *precision* dan *recall* yang tercapai masih rendah. Hal ini disebabkan tidak tersedianya cukup publikasi pada data latih untuk membentuk topik model yang representatif dari setiap peneliti. Hasil eksperimen juga menunjukkan bahwa CTL Berbasis Frase memiliki performa yang lebih baik daripada CTL berbasis kata. Terdapat peningkata nilai *precision* sebesar ±10% pada 10 rekomendasi teratas dan ±5% pada 20 rekomendasi teratas.

Referensi

Blei, D. M., Lafferty, J. D., 2009, Visualizing Topics with Multi-Word Expressions, 1–12. Retrieved from http://arxiv.org/abs/0907.1013

Danilevsky, M., Wang, C., Desai, N., Guo, J., Han, J., 2013, KERT: Automatic Extraction and Ranking of Topical Keyphrases from Content-Representative Document Titles. Retrieved from http://arxiv.org/abs/1306.0271

El-Kishky, A., Song, Y., Wang, C., R. Voss, C., Han, J., 2014, Scalable Topical Phrase Mining from Text Corpora. Proceedings of the VLDB Endowment

Fujiwara, Y., Nakatsuji, M., Onizuka, M., Kitsuregawa, M., 2012, Fast and Exact Top-k Search for Random Walk with Restart. In Proceedings of the VLDB Endowment, hal 442–453. https://doi.org/10.14778/2140436.2140441

Han, J., Wang, C., 2014, Mining latent entity structures from massive unstructured and interconnected data. Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14. https://doi.org/10.1145/2588555.2588890

Hemptinne, J. De, Ferrasse, J., Gorak, A., Kjelstrup, S., Maréchal, F., Baudouin, O., Gani, R., 2017, Chemical Engineering Research and Design Energy efficiency as an example of cross-discipline. Chemical Engineering Research and Design, 119, hal 183–187. https://doi.org/10.1016/j.cherd.2017.01.020

Kang, Y.B., Li, Y.-F., Coppel, R. L., 2015, Capturing Researcher Expertise through MeSH Classification. Proceedings of the Knowledge Capture Conference on ZZZ - K-CAP 2015, 1–8. https://doi.org/10.1145/2815833.2815837

Khamidah, F. S. N., Hapsari, D., Nugroho, H., 2018, Implementasi Fuzzy Decision Tree Untuk Prediksi Gagal Ginjal Konis. Integer: Journal of Information Technology, Vol 3, No 1, hal 19-28

Liang, W., Zhou, X., Huang, S., Hu, C., Xu, X., Jin, Q., 2017, Modeling of Cross-disciplinary Collaboration for Potential Field Discovery and Recommendation Based on Scholarly Big Data. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2017.12.038

Lindsey, R. V., III, W. P. H., Stipicevic, M. J., 2012, A Phrase-Discovering Topic Model Using Hierarchical Pitman-Yor Processes. Proceedings of the 2012 Joint Conerence on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, (July), hal 214–222.

Mitchell, M., Moore, S. A., Clement, S., Lockwood, M., Anderson, G., Gaynor, S. M., Lefroy, E. C., 2017, Biodiversity on the brink : Evaluating a transdisciplinary research collaboration. Journal for Nature Conservation, 40(December 2016), 1–11. https://doi.org/10.1016/j.jnc.2017.08.002

Osuna, F., Akbar, M., Gates, A. Q., 2017, On Using Disparate Scholarly Data to Identify Potential Members for Interdisciplinary Research Groups. 2017 IEEE International Conference on Information Reuse and Integration (IRI), 59–68. https://doi.org/10.1109/IRI.2017.33

Pangestu, B., Purwitasari, D., Fatichah, C., 2017, Visualisasi Similaritas Topik Penelitian dengan Pendekatan Kartografi Menggunakan Self Organizing Maps (SOM). Jurnal Teknik ITS, Vol 6 No 2, hal 417-420. http://dx.doi.org/10.12962/j23373539.v6i2.23706

Pay, T., 2016, Totally automated keyword extraction. Proceedings - 2016 IEEE International Conference on Big Data, Big Data 2016, hal 3859–3863. https://doi.org/10.1109/BigData.2016.7841059

Purwitasari, D., Fatichah, C., 2017, Inter-Departemental Research Collaboration Recommender System based on Content Filtering in a Cold Start Problem. 2017 IEEE 10th International Workshop on Computational Intelligence and Applications, hal 177-184. https://doi.org/10.1109/IWCIA.2017.8203581

Santoso, M., Indriyani, T., Putra, R.E., 2017., Deteksi Microaneurysms Pada Citra Retina Mata Menggunakan Matched Filter. Integer: Journal of Information Technology, Vol 2, No 2, hal 59-68

Schone, P., Jurafsky, D., 2001, Is knowledge-free induction of multiword unit dictionary headwords a solved problem? Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing, 100–108.

Tang, J., Wu, S., Sun, J., Su, H., 2012, Cross-domain collaboration recommendation. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-KDD '12, 1285. https://doi.org/10.1145/2339530.2339730

Wang, Q., Ma, J., Liao, X., Du, W., 2017, A context-aware researcher recommendation system for university-industry collaboration on R & D projects. Decision Support Systems, 103, hal 46–57. https://doi.org/10.1016/j.dss.2017.09.001

Wang, X., McCallum, A., Wei, X., 2007, Topical N-grams: Phrase and topic discovery, with an application to information retrieval. Proceedings - IEEE International Conference on Data Mining, ICDM, hal 697–702. https://doi.org/10.1109/ICDM.2007.86