pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 61 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL

HYPERGRAPH-PARTITIONING PADA *CO-AUTHORSHIP GRAPH* UNTUK PENGELOMPOKAN PENULIS BERDASARKAN TOPIK PENELITIAN

Daniel Swanjaya[1], Chastine Fatichah[2] , Diana Purwitasari [3]
[1]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Nusantara PGRI Kediri
swanjayadaniel@email.com
[2, 3]Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Surabaya
chastine.fatichah@email.com, diana.purwitasari@email.com

ABSTRAK

Topik Penelitian dapat diketahui dari Abstraksi dokumen penelitian, misalnya laporan Karya Tulis Ilmiah (KTI) berupa Tugas Akhir, Tesis dan Disertasi. Topik Penelitian dari KTI merupakan kumpulan kata-kata penting yang menunjukkan area/bidang penelitian dari KTI tersebut. Sebuah KTI dibimbing beberapa Dosen pembimbing, dan seorang Dosen biasanya akan membimbing beberapa topik tertentu. Beberapa Dosen yang memiliki bidang penelitian yang sama membentuk grup riset dalam lingkup Jurusan, tetapi beberapa Jurusan terdapat Dosen yang memiliki kesamaan atau kemiripan bidang penelitian. Pada Tesis ini diusulkan metode untuk mengelompokkan Penulis (Dosen) berdasarkan kesamaan topik penelitian pada Co-Authorship Graph menggunakan Hypergraph Partitioning, sehingga memungkinkan untuk membuat grup riset dalam lingkup antar Jurusan atau tingkat perguruan tinggi. Metode dibagi menjadi tiga tahap yaitu ekstraksi topik penelitian, pembentuksn *Co-Authorship Graph*, dan pengelompokan Penulis. Ekstraksi topik penelitian, mendapatkan topik dari KTI berdasarkan Judul dan Abstraksi menggunakan Latent Dirichlet Allocation (LDA). Pembentukan *Co-Authorship Graph*, dimana node adalah Penulis, edge adalah hubungan kolaborasi dan kesamaan/kemiripan topik penelitian, dan bobot edge adalah nilai jaccard dan cosine similary topik penelitian antar Penulis. Pengelompokkan Penulis pada *Co-Authorship Graph* menggunakan *Hypergraph Partitioning*. Uji coba metode menggunakan data Penelitian dari Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITS Surabaya. Hasil pengelompokkan divalidasi menggunakan *Silhouette dan Entropy*. Hasil akhir pengelompokkan menunjukkan bahwa telah terbentuk kelompok Penulis yang anggotanya berasal dari Jurusan atau bidang yang berbeda, dengan kesamaan topik yang tinggi.
Kata kunci: *Graph Clustering, Latent Dirichlet Allocation, Co-Authorship Graph, Hypergraph Partitioning*

Abstract

Research topics can be seen from Abstraction research documents, for example, reports Scientific Writing (KTI) in the form of Final Project, Thesis and Dissertation. Research Topics of KTI is a collection of important words that indicate the area / field of study of the KTI. A guided KTI some supervisor, and a lecturer normally would guide some particular topic. Some lecturers have the same field of research formed a research group within the Department, but some courses are lecturers who exhibit similarities field of research. At this thesis proposed a method for classifying Writer (Lecturer) based on common research topics in Co-Authorship Graph using the Hypergraph Partitioning, making it possible to create a research group within the scope of inter Programs or college level. The method is divided into three stages: extraction of research topics, pembentuksn Co-Authorship Graph, and grouping author. Extraction of research topics, get the topic of EI by Title and Abstract using Latent Dirichlet Allocation (LDA). Formation of Co-Authorship Graph, where the nodes are the author, edge is the collaborative relationship and similarity / resemblance of research topics, and the weighting edge is Jaccard and cosine values similary research topics between author. Grouping Writers on Co-Authorship Graph using the Hypergraph Partitioning. Test method uses data from the Research Institute of Research and Community Service (LPPM) ITS. Grouping the results are validated using the Silhouette and Entropy. The final results showed that the grouping has been formed group Authors whose members come from the Department or a different field, with high similarity topic.
Keywords: *Graph Clustering, Latent Dirichlet Allocation, Co-Authorship Graph, Hypergraph Partitioning*

I. PENDAHULUAN

Setiap karya tulis pasti memiliki topik pembicaraan, demikian pula dengan Karya Tulis Ilmiah (KTI) di perguruan tinggi yang memiliki minimal satu topik, yang umumnya disebut sebagai topik penelitian. Karya yang *term*asuk KTI adalah Skripsi/Tugas Akhir, Tesis dan Disertasi. Pada karya tulis, umumnya diberikan informasi kata kunci (*keyword*) untuk merepresentasikan kata penting

dalam karya tulis tersebut, tetapi informasi tersebut belum dapat dijadikan acuan sebagai topik penelitian, kata kunci umumnya hanya digunakan untuk membantu pencarian KTI pada sistem informasi perpustakaan. Untuk menemukan topik pada KTI pembaca harus membaca keseluruhan isi karya tulis, tetapi hal tersebut membutuhkan waktu yang lama. Selain kata kunci terdapat juga informasi Abstraksi yang berisi uraian singkat dari isi penelitian secara menyeluruh. Topik penelitian bisa dapat ditemukan pada Abstraksi, tetapi untuk menemukan topik penelitian karya tersebut pembaca juga harus membaca abstraksi sepenuhnya, padahal terkadang pembaca tidak mempunyai waktu yang singkat.

Informasi topik penelitian pada karya tulis dalam bentuk digital dapat diperoleh dengan beberapa metode. Penelitian [5] mengusulkan metode untuk identifikasi topik menggunakan *Hypergraph Partitioning*. Ekstraksi topik dilakukan untuk mengumpulkan kata-kata kunci yang ada pada suatu koleksi dokumen sehingga dapat digunakan untuk mengenali topik. Hubungan antar kata dalam koleksi dokumen dimodelkan menjadi suatu *Hypergraph* dengan kata-kata sebagai node dan kekuatan hubungan antar kata sebagai *edge* yang memiliki bobot (weighted *edge*). Partisi dilakukan dengan metode *Hypergraph Partitioning* dengan memotong *graph* yang ada menjadi sub-sub *graph* yang berisi kata-kata kunci (keyword) untuk mengenali topik. Uji coba dilakukan untuk mengukur tingkat ketepatan identifikasi topik menggunakan data set *standard 20 Usenet newsgroups* milik *UCI KDD Archive*, sejumlah 108 dokumen berbahasa Inggris dengan 4 kategori topik dan 8 kategori subtopik. Hasilnya pada tiap partisi dapat dikenali topik apa yang sedang dibicarakan. Penentuan topik dari suatu dokumen secara otomatis dapat dilakukan dengan metode ekstraksi topik, beberapa penelitian berita berbahasa Indonesia ataupun asing sebagai percobaannya.

Penelitian [8] menggunakan *Probabilistic Latent Semantic* Analysis (PLSA) untuk mengelompokkan kata-kata ke dalam topik-topik yang belum diketahui (*latent*), kemudian menggunakannya untuk mengklasterkan dokumen. [1] menggunakan Latent Semantic Analysis (LSA) dan Singular Value Decomposition (SVD) untuk mengekstraksi topik-topik utama harian dari kumpulan dokumen berita online berbahasa Indonesia.

Penelitian [9] mengusulkan metode untuk pengurutan kalimat berdasarkan topik kata kunci menggunakan LDA. Metode yang diusulkan memiliki tiga tahapan. Pertama, mengelompokkan kalimat-kalimat pada setiap dokumen menggunakan *similarity histogram clustering* (SHC). Kedua, merangking *cluster* yang terbentuk menggunakan *cluster importance*. Ketiga, menyusun kalimat representatif yang dipilih dan disusun berdasarkan indentifikasi topik menggunakan LDA. Pengujian metode menggunakan data DUC 2004 dan dianalisa menggunakan ROUGE-1 dan ROUGE-2. Kalimat ringkasan yang dihasilkan koheren (bertalian secara logis) sehingga waktu untuk membaca ringkasan lebih singkat.

Dosen pembimbing yang membimbing Tugas Akhir/Skripsi, Tesis atau Disertasi membimbing sesuai dengan keahlian mereka. Pada karya yang dihasilkan oleh Mahasiswa, Dosen Pembimbing juga dianggap sebagai Penulis. Beberapa karya tulis merupakan gabungan dari beberapa disiplin ilmu, sehingga melibatkan beberapa Dosen Pembimbing, kerjasama ini sering disebut sebagai kolaborasi. Kolaborasi yang terjalin ini memiliki konsistensi, dimana masing-masing memiliki pasangan yang tetap sesuai dengan bidang penelitian yang diminati. Dosen-dosen yang memiliki kesamaan/kemiripan keahlian dan sering berkolaborasi ini membentuk kelompok penelitian atau grup riset dalam lingkup Jurusan. Tetapi di beberapa Jurusan lain pada perguruan tinggi yang sama juga terdapat topik penelitian atau bidang riset yang sama atau mirip, sehingga memungkinkan untuk membentuk grup riset antar Jurusan.

Untuk menggambarkan hubungan antar Dosen dalam hal penulisan karya tulis, digunakan Graf dimana *node*-nya Penulis dan *edge*-nya adanya karya tulis yang pernah ditulis bersama, Graf ini disebut *Co-Authorship Graph*. Pada tahun 2008, penelitian [12] membuat *Co-Authorship Graph* yang disebut graf komunikasi, data yang digunakan adalah Jurnal Penelitian dan Pengembangan Pertanian (Jurnal Litbang Pertanian) serta *Indonesian Journal of Agricultural Science* (IJAS) tahun 2005-2006, dimana informasi yang digunakan adalah nama penulis dan makalah yang dihasilkan oleh minimal dua penulis. Pada hasil penelitian diketahui tingkat kolaborasi peneliti bidang pertanian dan peneliti yang sering berkolaborasi merupakan peneliti yang produktif dan merupakan titik sintetis bila dibandingkan dengan peneliti yang jarang atau tidak berkolaborasi, serta menunjukkan bahwa jaringan komunikasi antar peneliti melalui artikel ilmiah yang dipublikasikan pada Jurnal Litbang Pertanian dan IJAS tergolong tinggi/produktif.

pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 63 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL

Nhut T.H, dkk (2013) memanfaatkan *Co-Authorship Graph* untuk memprediksi topik dari sebuah makalah (*paper*). Mereka memiliki asumsi bahwa makalah yang bertetanggaan pada *Co-Authorship Graph* memiliki topik yang sama dan topik makalah yang akan dipresiksi bergantung pada topik-topik makalah yang terhubung dengan makalah tersebut. Dengan menggunakan data ILPnet2 yang berisi tentang informasi makalah dari ILP (*Inductive Logic Programming*) tahun 1970 sampai dengan 2003. Dari *Co-Authorship Graph* yang terbentuk diketahui adanya komunitas ilmiah atau grup riset dari penulis makalah tersebut, pasangan penulis yang produktif. Tetapi keberhasilan metode *Fast Algorithm* ini sangat dipengaruhi oleh tingkat kepadatan ketetanggaan pada *Co-Authorship Graph.*

Pengelompokkan Penulis pada *Co-Authorship Graph* dapat dilakukan dengan metode *clustering* yang ada, diantaranya Qi Y (2011) menganalisa dan mengekstrak grup riset dari *co-authorship network* pada *Oncology* di Cina. Dengan menggunakan centrality, component analysis, K-Core, M-Slice, Hierarchical Clustering analysis dan Multidimensional Scaling analysis. Pengujian menggunakan data dari 10 Core Chinese Oncology journals antara tahun 2000 sampai 2009. Tujuannya untuk menganalisa kerja sama grup riset pada *co-authorship network Chinese Oncology*, memilih kelompok penelitian yang paling produktif dan setiap individu dalam grup riset *Chinese Oncology*. Manfaat metode ini adalah memberikan saran kepada pembuat kebijakan untuk membangun sistem yang lebih efisien untuk mengelolan dan membiayai penelitian *Chinese Oncology* ke depannya.

Penelitian [2] mencari komunitas penulis dan hubungannya berdasarkan *co-authorship network*, menggunakan dua dataset, CiNii, informasi bibliografi tentang publikasi ilmiah Jepang sejak tahun 1886, dan DBLP, informasi bibliografi publikasi ilmiah di bidang Ilmu Komputer tahun 1986 sampai dengan 2012. Informasi yang ada pada masing-masing bibliografi adalah nama publikasi, nama penulis, nama jurnal, tanggal publikasi dan kutipannya. Dari data tersebut dibangun masing-masing *co-authorship network*-nya dengan tidak mengikutsertakan penulis yang memiliki jumlah karya kurang dari *t*. Kemudian menggunakan metode yang dikembangkan oleh [4] yaitu *heuristic method* berdasarkan pada *modularity optimization*, untuk menemukan komunitas penulis yang ada pada kedua *co-authorship network* tersebut. Hasilnya didapatkan bahwa aspek yang mempengaruhi karakteristik sebuah komunitas adalah total ukuran, diameter, radius, *density*, *average degree*, jumlah sisi dalam grup, jumlah sisi luar grup, rasio *InDegree-OutDegree*, identifikasi node yang paling penting berdasarkan degree-nya, mengidentifikasi bidang studi dan Identifikasi penulis.

Pengelompokkan Penulis yang ada sampai saat ini hanya berdasarkan kolaborasi Penulis, kesamaan atribut karya yang dibuat (*jaccard*), hasil pengelompokkan hanya beranggotakan Penulis yang pernah berkolaborasi saja, tetapi dari hasil tersebut terdapat juga beberapa Penulis yang memiliki kesamaan topik karya tulis, yang tersebar pada beberapa kelompok. Diharapkan apabila Penulis-Penulis yang memiliki kesamaan topik dapat dikelompokkan maka dapat dibuat grup riset atau grup Penulis baru sehingga dapat meningkatkan produktifitas Penulis dan mencegah terjadinya kesamaan karya tulis.

Topik Penelitian pada beberapa Jurusan atau Program Studi memiliki kemiripan atau kesamaan, sehingga memungkinkan untuk mengelompokkan Dosen yang memiliki topik penelitian yang mirip sehingga didapatkan grup penelitian yang merupakan kolaborasi antar jurusan. Pada Tesis ini diusulkan penggunaan *Hypergraph Partition* pada *Co-Authorship Graph* untuk mengelompokkan Penulis berdasarkan topik penelitian. Usulan ini dibagi menjadi tiga tahap, ektraksi topik penelitian, pembentukan *Co-Authorship Graph* dan pengelompokkan Penulis menggunakan *Hypergraph-Partitioning*. Pada tahap ekstraksi topik penelitian dilakukan proses pembersihan pada data Dokumen (Judul dan Abstraksi KTI) yang kemudian diekstraksi menggunakan LDA sehingga didapat probabilitas topik Dokumennya, kemudian dibuat representasi probabilitas topik dari tiap Penulis berdasarkan data Penulis Dokumen dan probabilitas topik dokumen. Pada tahap pembentukan *Co-Authorship Graph*, koefisien Jaccard antar Penulis, yang merepresentasikan kolaborasi atau kerjasama, dan similaritas topik penelitian antar Penulis digunakan untuk menentukan bobot hubungan antar Penulis, yang kemudian digunakan untuk membuat *Co-Authorship Graph* dimana *node*-nya adalah Penulis dan *edge*-nya adalah adanya kesamaan/kemiripan antar penulis. Pada tahap pengelompokkan Penulis, *Hypergraph-Partitioning multilevel* digunakan untuk mempartisi *node-node* pada *co-aouthorship graph*, sehingga terbentuk

kelompok-kelompok Penulis. Data yang digunakan pada penelitian ini adalah data Karya Tulis Ilmiah Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITS Surabaya sebanyak 10.722 KTI dari 22 Jurusan, dan evaluasi hasil pengelompokan menggunakan *entropy* dan *Silhouette coefficient*.

II. DATASET

Data yang digunakan pada penelitian ini adalah data Karya Tulis Ilmiah yang ada di Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITS Surabaya, yang terdiri dari lima Fakultas yaitu, Fakultas Matematika dan Ilmu Pengetahuan (FMIPA), Fakultas Teknologi Industri (FTI), Fakultas Teknik Sipil dan Perencanaan (FTSP), Fakultas Teknologi Kelautan (FTK), Fakultas Teknologi Informasi (FTIF). Informasi KTI yang digunakan adalah Judul karya tulis (*Title*), Dosen Pembimbing 1 (*Author_1*), Dosen Pembimbing 2 (*Author_2*), Dosen Pembimbing 3 (*Author_3*), Abstraksi karya tulis (*Abstact*).

Seperti yang ditunjukkan Tabel 1 dan Tabel 2, banyaknya karya tulis yang digunakan berjumlah 10.722 data yang berasal dari 22 Jurusan, dimana ada 6.809 karya tulis yang dibimbing oleh seorang dosen, 3.859 oleh 2 dosen dan 54 oleh 3 dosen.

III. DASAR TEORI

Pada bagian ini menjelaskan dasar teori yang berhubungan dengan penelitian yang diusulkan meliputi, *Latent Dirichlet Allocaton* dan *Hypergraph Partitioning*.

A. Latent Dirichlet Allocaton

Beberapa metode untuk mengekstrak topik adalah *Latent Semantic Analysis* (LSA), Probabilistic Latent Semantic Analysis (pLSA) dan *Latent Dirichlet Allocation* (LDA). Pada metode LSA saat mengekstraksi topik memperhatikan adanya sinonim (arti kata sama) dan polisemi (kata yang mempunyai banyak arti), hal ini menjadi kelemahan dari LSA karena harus membangun kamus sinonim dan polisemi terlebih dahulu. Metode pLSA adalah perkembangan dari LSA, dengan menambahkan model probabilistik, tetapi untuk menggunakan pLSA kita harus menyusun urutan dokumen dengan benar, apabila tertukar akan memberikan hasil yang berbeda. LDA menyempurnakan pLSA dengan membuang ketergantungan pada urutan dokumen. Permodelan LDA terdapat pada Gambar 1.

TABEL I
BANYAK KARYA TULIS PADA TIAP JURUSAN DI ITS SURABAYA

Jurusan	Banyak Karya Tulis
Fisika	270
Matematika	328
Statistika	615
Kimia	299
Biologi	151
Teknik Mesin	1.026
Teknik Elektro	1.544
Teknik Kimia	575
Teknik Fisika	558
Teknik Industri	581
Teknik Material Dan Metalurgi	200
Teknik Sipil	1.145
Arsitektur	423
Teknik Lingkungan	424
Desain Produk	347
Teknik Geomatika	159
Perencanaan Wilayah Kota	170
Teknik Perkapalan	282
Teknik Sistem Perkapalan	358
Teknik Kelautan	299
Teknik Informatika	639
Sistem Informasi	329
Total	10.722

Pada level pertama, terdapat parameter α dan β. Parameter α merupakan distribusi *Dirichlet* untuk distribusi toipik pada satu dokumen, secara umum nilainya adalah 50/K, dimana K adalah banyak topik. Parameter β merupakan distribusi *Dirichlet* untuk distribusi kata dalam satu topik. Parameter-parameter tersebut di-*sample* satu kali pada saat proses me-*generate corpus*. Pada level kedua, variabel θ menunjukkan distribusi masing-masing

pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 65 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL

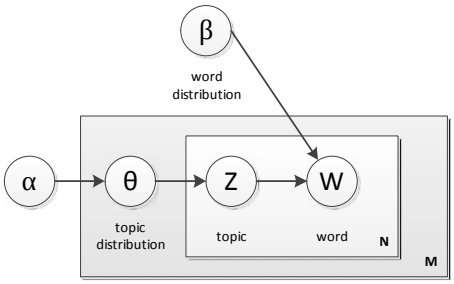
topik pada masing-masing dokumen. Pada level ketiga, variabel z merupakan topik-topik yang terdapat di dalam *corpus*, sedangkan variabel w merupakan kata-kata yang terdapat di dalam *corpus*. Variabel-variabel tersebut di-*sample* satu kali pada masing-masing kata dan masing-masing dokumen.

B. Hypergraph Partitioning

Proses membagi sebuah graf G menjadi beberapa himbpunan yang saling lepas. Jika graf G = {V, E} dan partisi P = {S_1, S_2, S_3, ... , S_n} dimana S adalah himpunan yang berisikan n partisi, maka representasi dari v terhadap p didefinisikan sebagai r(v|p) = (d(v,S_1), d(v,S_2), d(v,S_3), ... d(v,S_n)). Pada graf yang memiliki jumlah node banyak, tidak dapat dilakukan secara langsung karena akan mengakibatkan kompleksitas perhitungan akan sangat tinggi, oleh karena itu perlu dilakukan tiga tahap untuk melakukan proses partisi pada graf berukuran besar, seperti pada Gambar 2.

TABEL II
BANYAK KARYA TULIS BERDASARKAN BANYAKNYA PENULIS, TANPA MAHASISWA

Jumlah Penulis	Jumlah Karya
1	6.809
2	3.859
3	54
Total	10.722



Gambar 1. Permodelan LDA

C. Hypergraph Partitioning

Proses membagi sebuah graf G menjadi beberapa himbpunan yang saling lepas. Jika graf G = {V, E} dan partisi P = {S_1, S_2, S_3, ... , S_n} dimana S adalah himpunan yang berisikan n partisi, maka representasi dari v terhadap p didefinisikan sebagai r(v|p) = (d(v,S_1), d(v,S_2), d(v,S_3), ... d(v,S_n)). Pada graf yang memiliki jumlah node banyak, tidak dapat dilakukan secara langsung karena akan mengakibatkan kompleksitas perhitungan akan sangat tinggi, oleh karena itu perlu dilakukan tiga tahap untuk melakukan proses partisi pada graf berukuran besar, seperti pada Gambar 2.

Pertama, fase coarsening, Graf G disederhakan menjadi dua bagian menggunakan algoritma Maxima Matching, Kedua dilakukan proses Partisi menggunakan algoritma Fiduccia-Mattheyeses sehingga *edge* yang terpotong seminimal mungkin, Terakhir fase *uncoarsening* , pada tahap ini node hasil dari fase coarsening yang telah terkelompok dibuka kembali.

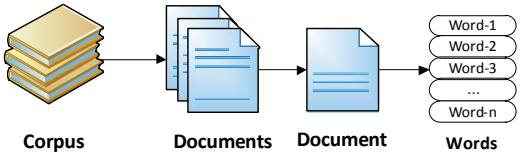
1. *Maxima Matching (Greedy)*

Algoritma *Maxima Matching* didefinisikan sebagai berikut:

- Untuk tiap node cari tetangga yang memiliki bobot paling tinggi, dimana node tetangga bukan node pasangan dari node lainnya.
- Kelompokkan node yang berpasangan.

Hitung bobot antar kelompok, bobot terbesar dari bobot antar node yang berpasangan.

- Lakukan hingga tidak ada kelompok yang bisa dikelompokkan atau terbentuk dua kelompok.



Gambar 2. Hubungan Corpus, Documents, Document dan Words.

2. Algoritma *Fiduccia-Mattheyeses*

Algoritma *Fiduccia-Mattheyeses*, FM [6] adalah algoritma *bisection Partitioning* yang bersifat *heuristic*. Algoritma ini menerapkan konsep menukar per-satu *node* pada setiap iterasinya. FM dimulai dengan *initial Partitioning*, *node – node* dengan algoritma Gredy dikelompokkan menjadi 2. Pada awal proses semua *node* bebas untuk bergerak (*unlocked*), dan setiap kemungkinan pergerakan ditandai dengan *gain*. Secara berulang, *node* yang memiliki *gain* terbesar dengan status unlocked akan dipindah posisi partisinya dan kemudian dikunci (locked), kemudian hitung ulang nilai *gain* semua *node*. Langkah-langkahnya :

a. Inisiasi semua node = unlocked.
b. Beri label pada kedua partisi, KIRI dan KANAN.
c. Selama ada node = unlocked, lakukan,
   1) Hitung nilai gain (Δg) tiap node pada partisi KIRI, untuk setiap node yang unlocked.

2) Cari node yang memiliki nilai gain terbesar pada partisi pertama, misal α.
3) Pindahkan α ke partisi KANAN.
4) α = locked.
5) Hitung nilai gain tiap node pada pastisi KANAN, untuk setiap node yang unlocked.
6) Cari node yang memiliki nilai gain terbesar pada partisi kedua, misal β.
7) Pindahkan β ke partisi KIRI.
8) β = locked.
9) Hitung jumlah bobot dari *edge* yang terpotong, akibat perpindahan α dan β.

d. Solusi adalah kondisi saat α dan β yang memiliki jumlah bobot dari *edge* yang terpotong paling besar ditukar.
Dimana :

Nilai gain tiap node. $\Delta g = FS - TE$.

FS(α) = jumlah bobot dari *edge* yang menghubungkan node α dengan node-node yang ada di luar partisi dimana node α berada.

TE(α) = jumlah bobot dari *edge* yang menghubungkan node α dengan node-node yang ada di dalam partisi dimana node α berada.

3. *Uncoarsening*

Pada tahap coarsening setahap demi setahap dilakukan penggabungan hingga terbentuk dua kelompok atau partisi, setelah dilakukan *Balancing* dengan Algoritma *Fiduccia-Mattheyeses*, dilakukan *uncoarsening* yaitu kebalikan dari coarsening, yaitu melakukan penyesuaian akibat dari proses *Balancing*, pertukaran node, untuk semua pengelompokkan pada tahap coarsening, setelah proses penyesuaian selesai dilakukan, maka dilakukan proses *Balancing* untuk child dari kelompok induknya.

IV. METODOLOGI

Pada penelitian ini diusulkan penggunaan *Hypergraph Partition* pada *Co-Authorship Graph* untuk mengelompokkan Penulis berdasarkan topik penelitian. Usulan ini dibagi menjadi tiga tahap, ektraksi topik penelitian, pembentukan *Co-Authorship Graph* dan pengelompokkan Penulis menggunakan *Hypergraph-Partitioning*.

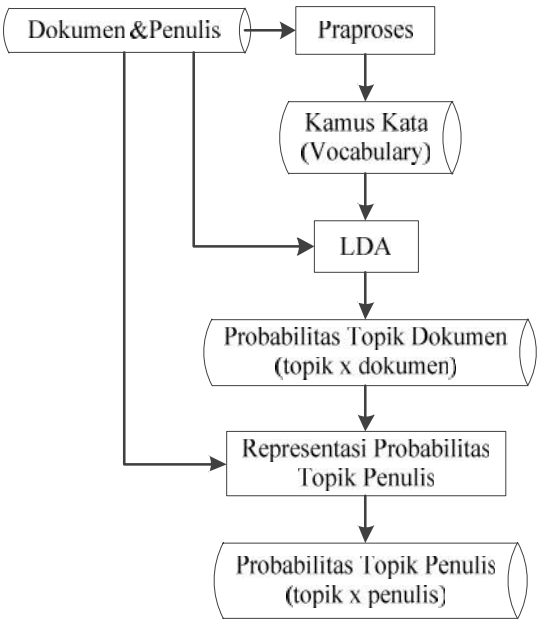
1. Ekstraksi topik penelitian dengan LDA

Pada praproses seperti yang ditunjukan Gambar 3, Dokumen yang berupa teks dibersihkan dengan cara *Case folding*, *Filtering*, *Stemming*, *Stoplist removal*, *Terms extraction*, *Weighting*.

Langkah berikutnya adalah ekstraksi topik dari Dokumen-Dokumen yang telah dibersihkan menggunakan *Latent Dirichlet Allocation* (LDA), sehingga dihasilkan probabilitas setiap *term* terhadap topik dan probabilitas setiap topik terhadap Dokumen. Dari probabilitas setiap *term* terhadap topik, didapatkan *term* yang dominan pada tiap topik, dan dari probabilitas setiap topik terhadap Dokumen didapat Topik dominan dari tiap Dokumen. Berdasarkan informasi Penulis Dokumen dan probabilitas setiap topik terhadap Dokumen, didapatkan representasi probabilitas topik terhadap Penulis. Probabilitas topik dari tiap Penulis didapatkan dari jumlah probabilitas topik tiap Dokumen yang ditulis oleh Penulis tersebut, kemudian jumlah probabilitas topik dinormalisasi sehingga total dari probabilitas tiap topik dari seorang Penulis adalah satu.

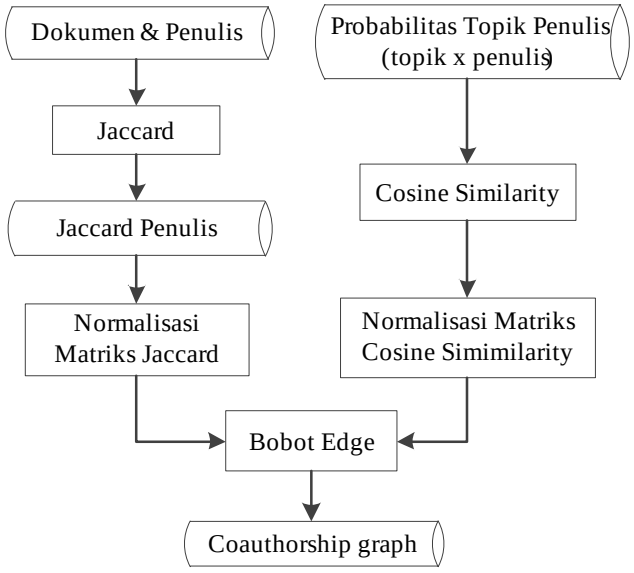
2. Pembentukan Co-Authorship Graph

Co-Authorship Graph digunakan untuk menggambarkan hubungan antar Penulis, dimana Penulis sebagai *node* dan hubungan antar penulis sebagai *edge*. Bobot *edge* ditentukan dari nilai kesamaan (*cosine similarity*) antar topik penelitian Penulis dan koefisien Jaccard kerjasama antar Penulis. Koefisien *jaccard* penulis adalah nilai perbandingan jumlah karya tulis hasil kolaborasi kedua Penulis dengan total jumlah karya tulis masing-masing penulis. Pada persamaan 1 formula *jaccard* dimodifikasi dengan menambahkan konstanta, 1, untuk menghindari adanya nilai 0 akibat dari tidak adanya kerjasama antar Penulis. Kesamaan topik dari kedua penulis didapatkan dari nilai *cosine similarity* dari *vector* probabilitas topik terhadap Penulis dari masing-masing Penulis.



Gambar 3. Proses ekstraksi Topik Penelitian

pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 67 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL



Gambar 4. Pembentukan *Co-Authorship Graph*.

$$Jaccard(A,B) = \left(\frac{KTI\ (A\ dan\ B)}{KTI\ (A) + KTI\ (B) - KTI\ (A\ dan\ B)}\right) \quad (1)$$

$$+ 1$$

$$+ 1 \quad (2)$$

$$Similarity(A,B) = \frac{A.B}{||A||||B||}$$

$$Bobot(A,B) = c \times Similarity\ (A,B) + (1-c) \times Jaccard(A,B) \quad (3)$$

Sebelum menentukan bobot *edge*, nilai Similarity dan Jaccard dinormalisasi terlebih dahulu, sehinnga nilainya mempunyai jangkauan 0 hingga 1. Kemudian dipilih sebuah nilai konstanta, *c*, sebagai nilai perbandingan komposisi antara kesamaan topik dengan kerjasama. Dengan menggunakan nilai konstanta, *c*, similaritas dan jaccard yang sudah dinormalisasi didapatkan nilai bobot *edge* yang merepresentasikan nilai relasi antar Penulis, seperti pada Persamaan 3. Pembentukan Co-Authorship Graph terdapat pada Gambar 4.

3. Pengelompokkan menggunakan Hypergraph-Partitioning

Proses pengelompokkan Penulis terdiri dari 3 tahap. Coarsening, inisiasi partisi dimana semua node dikelompokkan menjadi 2 kelompok atau lebih jika pada saat pengelompokkan terdapat beberapa kelompok yang tidak memiliki hubungan. *Balancing* menggunakan algoritma Fiduccia-Mattheyses, meminimalis banyak *edge* yang terpotong setelah dilakukan inisiasi partisi. *Uncoarsening* , dengan merujuk pada tahapan Coarsening dilakukan *Balancing* pada setiap pasangan partisi dengan penyesuaian dari proses *Balancing* dari partisi sebelumnya. Setelah tahap *Uncoarsening* selesai, berikutnya menentukan maksimal banyak anggota dari tiap kelompok, untuk mendapatkan kelompok Penulis. Proses ini ditunjukan pada Gambar 5.

V. HASIL DAN PEMBAHASAN

Pengujian penelitian ini dilakukan dengan tiga pengujian dan dievaluasi menggunakan Entropy dan *Silhouette* Coefficient.

A. Entropy

Entropy adalah suatu parameter yang menunjukkan tingkat kemurnian dari klaster yang terbentuk. Entropy (Zhao, 2001) dihitung berdasarkan matriks confusion hasil klasterisasi dalam persamaan :

$$E(C^i) = -\frac{1}{\log q}\sum_{j=1}^{q}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i} \quad (4)$$

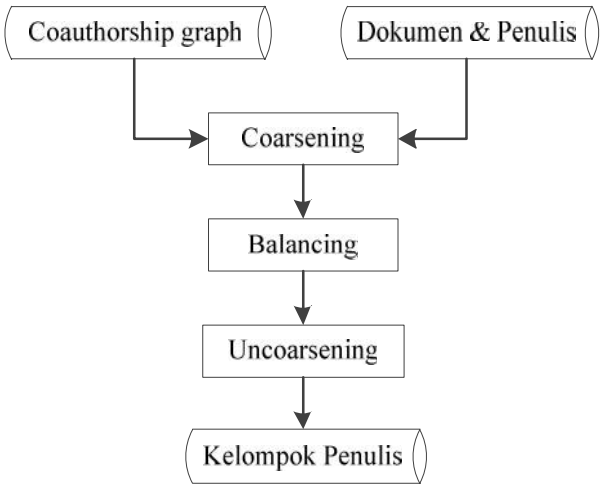
$$E(C_i) = -\frac{1}{\log q}\sum_{j=1}^{q}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i}$$

$$Entropy = \sum_{i=1}^{k}\frac{n_i}{n}E(C^i) \quad (5)$$

Dimana,

: Nilai Entropy Cluster ke-i
: Banyak class pada GroudTruth.
: Banyak anggota *cluster* ke-i yang *term*asuk class ke-j
: Banyak anggota *cluster* ke-i
: Banyak seluruh anggota.
: Banyak *cluster*.

Semakin tinggi nilai Entropy, maka semakin baik kualitas kelompok yang terbentuk



Gambar 5. Pengelompokkan Penulis menggunakan *Hypergraph-Partitioning*

B. *Silhouette*

Silhouette coefficient mengkombinasikan ide cohesion dan separation untuk validasi hasil klastering. *Cohesion* digunakan untuk mengukur seberapa dekat hubungan objek-objek pada klaster yang sama. Sedangkan *separation* digunakan untuk mengukur seberapa berbeda atau terpisahnya sebuah klaster dari klaster lainnya. Sedangkan *Silhouette coefficient* sendiri digunakan untuk mengukur kualitas klaster yang dihasilkan sekaligus mengindikasikan derajat kepemilikan setiap objek yang berada di dalam klaster. Nilai *Silhouette* setiap objek dengan persamaan :

Nilai *Silhouette* akan mengindikasikan derajat kepemilikan tiap objek ditampilkan pada Tabel 3.

$$b(x) = \max_{Ci \neq A} d(x, C^i) \quad (6)$$

$$S_i = \frac{a(x) - b(x)}{\max(a(x), b(x))} \quad (7)$$

TABEL III
INTEPRESTASI NILAI SILHOUETTE OBYEK

S(x)	Intepretasi
Negatif	Menunjukkan overlapping struktur yang tinggi, bahwa x berada dekat dengan objek lain di klaster B, bukan A, klaster sebelumnya. Atau bisa dikatakan x seharusnya tidak berada di dalam klaster A.
0	Menunjukkan x adalah irisan dari klaster A dan B.
Positif	Menunjukkan x memang milik klaster A.

C. *Uji coba 1 : Pengaruh konstanta, c, terhadap model graf*

Pada uji coba ini, pengaruh nilai *c* pada perhitungan bobot edge, persamaan 3, yang digunakan untuk membentuk *Co-Authorship Graph*, terhadap hasil pengelompokkan Penulis menggunakan *Hypergraph-Partitioning*. Data yang digunakan untuk pengujian ini adalah Dokumen Jurusan Teknik Informatika, FTIf, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 639 Dokumen dan 35 Penulis. Pada ekstraksi topik Dokumen digunakan K = 12 Topik, α = 2 dan β = 0,05.

TABEL IV
NILAI AVERAGE SILHOUETTE WIDTH, PADA UJI COBA 1.

c	*Average Silhouette Width*
0,1	0,064
0,2	0,099
0,3	0,113
0,4	0,144
0,5	0,203
0,6	0,106
0,7	0,117
0,8	0,113
0,9	0,129

Nilai *c* yang digunakan untuk menghitung bobot edge adalah 0,1 sampai 0,9, masing-masing bobot digunakan untuk membentuk graf. Setiap graf yang terbentuk dikelompokkan menggunakan *Hypergraph-Partitioning*, dimana pada uji coba ini banyak anggota *cluster* maksimalnya adalah 8 Penulis, dan banyak *cluster* yang dihasilkan sebanyak 5 kelompok. Tabel 4 merupakan nilai ASW dari pengelompokkan graf yang dibentuk dari

Hasil pengelompokkan Penulis menggunakan *Hypergraph-Partitioning* pada ujicoba ini menunjukkan bahwa bobot yang menggunakan nilai *c* = 0,5 menghasilkan nilai rata-rata *Silhouette* terbesar yaitu sebesar 0,203, dan *c* = 0,1 menghasilkan nilai rata-rata *Silhouette* terkecil sebesar 0,064.

Pada hasil pengelompokkan dengan *c* = 0,5, Cluster ke-1 mempunyai nilai *Silhouette Cluster* tertinggi sebesar 0,3, hal ini menunjukkan bahwa antara Cluster ke-1 dengan *cluster* lainnya terpisah dengan baik (*well-separated*). Anggota Cluster ke-1 masing-masing mempunyai nilai positif, hal ini mengindikasikan bahwa seluruh anggotanya terkelompok dengan baik. Berdasarkan nilai *Silhouette* anggota, pak Arya mempunyai nilai tertinggi, hal ini mengungkapkan bahwa tingkat persamaan pak Arya dengan ketujuh Penulis lainnya

pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 69 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL

sangat baik. Nilai *Silhouette* dari pak Yudhi adalah yang terkecil pada Cluster ke-1, hal ini mengungkapkan bahwa tingkat persamaan pak Yudhi dengan ketujuh Penulis lainnya kurang baik.

Cluster ke-4 mempunyai nilai *Silhouette* Width terkecil, hal ini mengindikasikan bahwa diantara kelima *cluster*, Cluster ke-4 memiliki keterpisahan yang paling kecil dengan *cluster* lainnya. Berdasarkan urutan nilai *Silhouette* anggota Cluster ke-4, didapatkan hubungan diantara keempat Penulis.

a. Suhadi Lili (0,058)
b. Imam Kuswardayan (0,055)
c. Dwi Sunaryono (0,031)
d. Riyanarto Sarno (0,030)

Keterkaitan antar anggota *cluster* kurang baik karena mendekati nol, pada *cluster* ini pak Riyanarto memiliki nilai terkecil, bahkan terkecil diantara semua Penulis, hal ini mengindikasikan bahwa tingkat kesamaan pak Riyanarto dengan semua penulis tidak sigifikan terutama dengan pak Suhadi, pak Imam dan pak Dwi.

Pengelompokkan menggunakan c = 0,1 menghasilkan pengelompokkan dengan nilai rata-rata *Silhouette* terkecil sebesar 0,064. Pada pengelompokkan ini terdapat satu Penulis dengan nilai *Silhouette* negatif, pak Fajar Baskoro, pada Cluster ke-2. Hal ini mengindikasikan bahwa pak Fajar tidak seharusnya berada di Cluster ke-2 karena nilai negatif berarti *intra-clusters similarity* lebih kecil daripada *inter-clusters similarity*-nya.

Pada pengujian ini nilai rata-rata *Silhouette* berkisar antara 0,064 sampai 0,203, maka berdasarkan Rousseeuw (1987), hasil pengelompokkan Penulis dengan *Hypergraph-Partitioning* masih "*No substantial structure has been found*" yang berarti *cluster* yang terbentuk tidak memiliki struktur yang alami.

Berdasarkan pengujian ini, ternyata nilai konstanta, *c*, pada perhitungan bobot memberikan pengaruh terhadap nilai rata-rata *Silhouette*, nilai terbesarnya diperoleh saat nilai konstanta, *c*, bernilai 0,5 atau seimbang, dan semakin kecil jika nilai *c*-nya semakin besar atau semakin kecil.

Sedangkan pengaruh variasi nilai *c* pada perhitungan bobot memberikan pengaruh pada pengelompokkan, nilai *c* yang kurang dari 0,5 akan meminimalisir kemungkinan Penulis mendapatkan nilai *Silhouette* negatif, pada nilai c = 0,1 satu Penulis, 0,2 tidak ada, 0,3 satu Penulis, dan 0,4 satu Penulis. nilai *c* yang lebih dari 0,5 akan memberikan nilai *Silhouette* negatif lebih banyak daripada nilai *c* yang kurang dari 0,5, pada nilai c = 0,6 empat Penulis, nilai c = 0,7 empat Penulis, nilai c = 0,8 empat Penulis, dan 0,9 tiga Penulis.

Pada Penelitian ini didapat kesimpulan bahwa model graf yang terbaik diperoleh dengan menggunakan nilai c = 0,5 atau dengan kata lain komposisi bobot kolaborasi dengan bobot kesamaan topik harus seimbang, sehingga menghasilkan hasil pengelompokkan yang baik.

D. *Uji coba 2 : Pengaruh konstanta, c, perhitungan bobot edge terhadap hasil pengelompokkan*

Pada uji coba ini, pengaruh nilai konstanta, *c*, pada perhitungan bobot, persamaan 3, terhadap hasil pengelompokkan akan diuji. Pengujian dilakukan dengan mengelompokkan Penulis dengan *Hypergraph Partitioning* berdasarkan bobot *edge* dari satu variasi nilai *c*.

Pengujian ini membuktikan bahwa dengan nilai *c* yang tinggi, bobot untuk kesamaan topik lebih besar dari bobot kolaborasi, akan menghasilkan kelompok Penulis baru.

Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Teknik Informatika saja, sebanyak 639 Dokumen dan 35 Penulis. dan pada ekstraksi topik Dokumen digunakan 12 Topik, $\alpha = 2$ dan $\beta = 0{,}05$.

Pada ujicoba ini akan dibandingkan hasil pengelompokkan antara bobot yang menggunakan c = 0,1 dan c = 0,9. Pada pengujian ini didapatkan kesimpulan bahwa hasil pengelompokkan dari graf yang menggunakan bobot kesamaan topik yang lebih besar dari bobot kolaborasi, memberikan kelompok baru dimana kesamaan topik anggotanya sangat baik. Tetapi terdapat kelompok yang anggotanya sama antara yang menggunakan bobot dengan nilai *c* tinggi dengan nilai *c* rendah, hal ini disebabkan dari tingginya kolaborasi antar Penulis dan Dokumen yang dihasilkan memiliki topik dominan yang tetap.

E. *Uji coba 3 : Perbandingan Hasil Pengelompokkan*

Pada Uji Coba ini, hasil dari pengelompokkan dengan *Hypergraph-Partitioning* akan dibandingkan dengan kelompok Peneliti dari data LPPM ITS dan divalidasi menggunakan Entropy.

Pengujian ini bertujuan untuk membuktikan asumsi bahwa dengan mengelompokkan Penulis berdasarkan kesamaan/kemiripan topik memungkinkan adanya kolaborasi baru antar Jurusan.

Data Dokumen yang digunakan adalah Dokumen karya tulis di lingkungan ITS Surabaya berupa Skripsi, Tesis dan Disertasi dari 5 Fakultas yang terdiri dari 22 Jurusan seperti pada Tabel 1, dan 751 Penulis. Pada tahap ekstraksi fitur digunakan K = 400 Topik, $\alpha = 2$ dan $\beta = 0{,}05$, kemudian pada tahap pengelompokkan Penulis maksimal banyak anggota *cluster* adalah 8, dan menghasilkan 94 *cluster*. Hasil pengelompokkan yang digunakan pada pengujian ini adalah Pengelompokkan yang memiliki nilai AWS terbesar. Pada Tabel 5, nilai rata-rata *Silhouette* yang tertinggi dicapai saat $c = 0{,}9$ sebesar 0,222.

TABEL V
NILAI RATA-RATA SILHOUETTE, PADA UJI COBA 3.

c	*Silhouette*
0,1	0,030
0,2	0,033
0,3	0,035
0,4	0,036
0,5	0,049
0,6	0,053
0,7	0,082
0,8	0,166
0,9	0,222

VI. KESIMPULAN

Term dominan pada Probabilitas kata terhadap topik, mampu merepresentasikan Topik dengan baik, dan Topik dominan dari probabilitas topik terhadap Dokumen mampu merepresentasikan Topik dari Penulis Dokumen. Nilai konstanta, *c*, pada perhitungan bobot edge dari *co-authorship graph* mempengaruhi hasil pengelompokkan, dimana nilai *Average Silhouette Width* (ASW) terbaik diperoleh saat $c = 0{,}5$. Hasil pengelompokkan dari Graf yang dibentuk dengan nilai *c* tertinggi ($c = 0{,}9$) mampu membentuk kelompok yang anggotanya memiliki kesamaan topik. Pembentukan kelompok Peneliti yang ada di Lingkungan ITS Surabaya masih berdasarkan pada pertemanan pada lingkup Jurusan, dan hanya sedikit kelompok yang terbentuk karena kesamaan topik. Hasil Pengelompokkan Penulis dengan lingkup lebih luas, pada penelitian ini digunakan data Dosen ITS Surabaya, mampu membentuk kelompok Penulis dengan anggota yang berasal dari beberapa Jurusan dengan kesamaan/kemiripan topik yang tinggi.

DAFTAR PUSTAKA

[1] Ashari N., "Ekstraksi Topik Utama Harian pada Portal Berita Indonesia Online menggunakan Singular Value Decomposition". *Skripsi Fakultas Matematika dan Ilmu Pengetahuan Alam*, Universitas Indonesia (2012).

[2] Bento, Carolina, and Hideaki Takeda. "Finding Research Communities and their Relationships by Analyzing the Co-authorship Network." *Information Visualisation* (IV), 2013 17th International Conference. IEEE, 2013.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022.

[4] Blondel, V.D., et al. "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008): P10008.

[5] Diana P, and Gestyana A.R. "Identifikasi Topik pada Koleksi Dokumen Menggunakan Algoritma Pengklasteran *Hypergraph Partitioning*." *Konferensi Nasional Sistem dan Informatika* 2011; Bali, November 12, 2011. 381-386.

[6] Fiduccia, Charles M., and Robert M. Mattheyses. "A linear-time heuristic for improving network *Partitions*." *Design Automation, 1982. 19th Conference on*. IEEE, 1982.

[7] Hoang, N.T, Phuc D, and Hoang N.L. "A Fast Algorithm for Predicting Topics of Scientific Papers Based on *Co-Authorship Graph* Model." *Advanced Methods for Computational Collective Intelligence*. Springer Berlin Heidelberg, 2013. 83-91.

[8] Ida A.G.S.P., Diana P., and Daniel O.S.. "Implementasi Algoritma Probabilistic Latent Semantic Analysis Dalam Pengklasteran Dokumen Berbasis Topik". Tugas Akhir Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember Surabaya (2011).

[9] Indra L, Daniel S, Arrie K and Agus ZA. "Multidocument Summarization Based on Sentence Clustering Improved Using Topic Words", *Jurnal Ilmiah Teknologi Informasi*. Vol. 12, No. 2, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Agustus 2014.

[10] Karypis, G, et al. "Multilevel *Hypergraph Partitioning*: applications in VLSI domain." *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*7.1 (1999): 69-79.

[11] Papa, David A., and Igor L. Markov. "*Hypergraph Partitioning* and

pISSN: 2442-3386 eISSN: 2442-4293
Vol 1 No 1 January 2015, 71 - 72
Melek IT
INFORMATION TECHNOLOGY JOURNAL

*cluster*ing."*Approximation algorithms and metaheuristics* (2007): 61-1.

[12] Vivit W.R. "Kolaborasi dan Graf Komunikasi Artikel Ilmiah Peneliti Bidang Pertanian: Studi Kasus pada Jurnal Penelitian dan Pengembangan Pertanian serta Indonesian Journal of Agricultural Science." *Jurnal Perpustakaan Pertanian* Vol. 17, Nomor 1, 2008

[13] Xiao, H, and Thomas Stibor. "Efficient Collapsed Gibbs Sampling for Latent Dirichlet Allocation." *Journal of Machine Learning Research-Proceedings Track*13 (2010): 63-78.

[14] Yu, Q., Hongfang S., and Zhiguang D. "Research groups of oncology co-authorship network in China." *Scientometrics* 89.2 (2011): 553-567

Halaman ini kosong
Redaksi Melek IT