TESIS – KI142502

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

Daniel Swanjaya
5112201022

DOSEN PEMBIMBING
Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
Diana Purwitasari, S.Kom., M.Sc.

PROGRAM MAGISTER
BIDANG KEAHLIAN KOMPUTASI CERDAS & VISUALISASI
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015

THESIS – KI142502

HYPERGRAPH-PARTITIONING ON *CO-AUTHORSHIP GRAPH* FOR AUTHOR CLUSTERING BASED ON RESEARCH TOPICS

Daniel Swanjaya
5112201022

Supervisor
Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
Diana Purwitasari, S.Kom., M.Sc.

MASTER DEGREE
COMPUTATIONAL INTELLIGENCE AND VISUALIZATION
INFORMATIC ENGINEERING DEPARTMENT
FACULTY OF INFORMATION TECHNOLOGY
SEPULUH NOPEMBER INSTITUTE TECHNOLOGY
SURABAYA
2015

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar

Magister Komputer (M.Kom.)

di

Institut Teknologi Sepuluh Nopember Surabaya

oleh :

Daniel Swanjaya

Nrp. 5112201022

Dengan judul :

Hypergraph-Partitioning pada *Co-Authorship Graph* untuk Pengelompokan Penulis Berdasarkan Topik Penelitian

Tanggal Ujian : 21 Januari 2015

Periode Wisuda : Maret 2015

Disetujui oleh :

Dr. Eng. Chastine Fatichah, S.Kom., M.Kom. NIP. 197512202001122002	..... (Pembimbing 1)
Diana Purwitasari,S.Kom., M.Sc. NIP. 197804102003122001	..... (Pembimbing 2)
Dr. Darlis Heru Murti,S.Kom. M.Kom. NIP. 197712172003121001	..... (Penguji 1)
Wijayanti Nurul Khotimah, S.Kom., M.Sc. NIP. 198603122012122004	..... (Penguji 2)
Bilqis Amaliah, S.Kom., M.Kom. NIP. 197509172001122002	..... (Penguji 3)

Direktur Program Pasca Sarjana,

Prof. Dr. Ir. Adi Soeprijanto, M.T.

NIP : 19640405 199002 1 001

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

Nama mahasiswa : Daniel Swanjaya
NRP : 5112201022
Pembimbing I : Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
Pembimbing II : Diana Purwitasari, S.Kom., M.Sc.

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 empat tahap yaitu praproses, ekstraksi topik penelitian, pembentukan *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.*

[Halaman ini sengaja dikosongkan]

HYPERGRAPH-PARTITIONING ON *CO-AUTHORSHIP GRAPH* FOR AUTHOR CLUSTERING BASED ON RESEARCH TOPICS

Nama mahasiswa : Daniel Swanjaya
NRP : 5112201022
Pembimbing I : Dr. Eng. Chastine Fatichah, S.Kom., M.Kom.
Pembimbing II : Diana Purwitasari, S.Kom., M.Sc.

ABSTRAK

Research topics are able to be known from the abstract section of research document, for example the document of literature, final project, thesis and dissertation. Research topics from literature are group of important word which are referred to the field of its literature. A literature is guided by some advisors and an advisor will handle some research topics. Advisors who have same research topics will form a research group in the faculty. In this thesis, a method is proposed to form group of researchers based on their topics with *Co-Authorship Graph*, it uses *Hypergraph-Partitioning* so it will possible to form research group in faculty or university. The method is divided to four phases, pre-processing, extracting research topic, forming *Co-Authorship Graph* and dividing authors. Research topic extraction, the topic is obtained from the title and the abstract using Latent Dirichlet Allocation (LDA). *Co-Authorship Graph* is formed, the node is author, the edge is collaboration and similarity research topic, and node's weight is jaccard value and cosine similarity of researcher topic. The authors agglomeration in *Co-Authorship Graph* use Hypergraph Partitioning. The testing method use the data from Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITS Surabaya. The result is validated by silhouette and entropy and it shows the agglomeration of authors. The member is obtained from the faculty or different field with high similarity topics.

Kata kunci: *Graph Clustering, Latent Dirichlet Allocation, Co-Authorship Graph, Hypergraph Partitioning.*

[Halaman ini sengaja dikosongkan]

KATA PENGANTAR

Segala puji dan syukur penulis ucapkan kepada Tuhan Yesus Kristus yang telah melimpahkan kasih, rahmat, anugerah dan karunia-Nya, sehingga penulis akhirnya dapat menyelesaikan Tesis dengan judul ***"HYPERGRAPH-PARTITIONING PADA CO-AUTHORSHIP GRAPH UNTUK PENGELOMPOKAN PENULIS BERDASARKAN TOPIK PENELITIAN "***,dapat terselesaikan dengan baik. Semoga tesis ini dapat memberikan manfaat bagi perkembangan ilmu pengetahuan terutama bidang komputasi cerdas dan visualisasi dan dapat memebri kontribusi bagi penelitian selanjutnya.

Penulis mengucapkan terima kasih atas bantuan dan dukungan dari berbagai pihak baik moril maupun materiil dalam pembuatan tesis ini, antara lain:

1. Dr. Eng. Chastine Fatichah, S.Kom., M.Kom. selaku dosen pembimbing utama atas kesabaran membimbing dan dukungan yang diberikan hingga terselesaikannya tesis ini.
2. Diana Purwitasari S.Kom., M.Sc. selaku dosen pembimbing kedua atas kesabaran membimbing dan dukungan yang diberikan hingga terselesaikannya tesis ini.
3. Dr. H. Agus Zainal Arifin, S.Kom, M.Kom, selaku Dekan Fakultas Teknologi Informasi ITS. Waskitho Wibisono, S.Kom., M.Eng., Ph.D., selaku Koordinator S2 Jurusan Teknik Informatika ITS.
4. Tim Penguji Tesis, Dr. Darlis Heru Murti, S.Kom., M.Kom., Wijayanti Nurul Khotimah, S.Kom., M.Sc., dan Bilqis Amaliah, S.Kom., M.Kom., selaku penguji siding tesis yang telah memberikan masukan dan arahan.
5. Bapak dan Ibu dosen pascasarjana Teknik Informatika ITS yang telah bersedia dengan sabar mengajar dan memberi bimbingan selama masa kuliah.
6. Keluarga besar Nobel atas segala doa, bimbingan, kasih sayang, perhatian, semangat, kerja keras, dan pengorbanannya selama ini.
7. Keluarga besar Universitas Nusantara PGRI Kediri tempat penulis mengajar dan mengembangkan kemampuan yang penulis miliki.
8. Keluarga besar Mahasiswa Pascasarjana Teknik Informatika Angkatan 2012, yang banyak memberikan inspirasi dan semangat selama kuliah serta dalam menyelesaikan Tesis ini.

9. Mbak Rini dan Mbak Veni atas bantuan informasi dan adminstrasi yang berkaitan dengan ujian proposal, dan ujian Tesis.
10. Mas Kunto atas kesediaan menjaga Laboratorium Pascasarjana Informatika ITS, sebagai tempat untuk penulis mengerjakan Tesis.
11. Semua teman-teman yang tidak tersebutkan, penulis mengucapkan terima kasih atas bantuannya.

Penulis menyadari bahwa Tesis ini masih jauh dari sempurna, sehingga kritik dan saran dari pembaca sangat penulis harapkan. Akhir kata, Akhir kata, penulis berharap semoga Penelitian ini dapat bermanfaat bagi banyak pihak terutama untuk pengembangan ilmu pengetahuan dan teknologi di bidang Komputasi Cerdas dan Visualisasi.

Surabaya, Januari 2015

Penulis

DAFTAR ISI

DAFTAR GAMBAR

[Halaman ini sengaja dikosongkan]

DAFTAR TABEL

BIOGRAFI PENULIS

Daniel Swanjaya lahir di Kediri pada tanggal 23 September 1983. Pada tahun 2008 Penulis telah menyelesaikan studi S1 sebagai Sarjana Komputer (S.Kom) di Jurusan Teknik Informatika, Universitas 17 Agustus 1945 Surabaya. Setelah itu Penulis mengajar di beberapa kampus di kota Kediri. Saat ini Penulis berstatus sebagai dosen tetap di Program Studi Teknik Informatika Universitas Nusantara PGRI Kediri (UNP Kediri). Pada tahun 2012 Penulis mendapat kesempatan untuk melanjutkan studi S2 di Program Pascasarjana Teknik Informatika ITS Surabaya dengan beasiswa dari DIKTI yaitu BPPS (Beasiswa Program Pasca Sarjana) atau BPP-DN (Beasiswa Pendidikan Pascasarjana Dalam Negeri). Pada Januari 2015 Penulis telah telah mengikuti ujian Tesis sebagai syarat mendapatkan gelar Magister Komputer di Institut Teknologi Sepuluh Nopember (ITS) Surabaya. Penulis memiliki minat riset di bidang *Dokumen processing* dan *text mining*.

Email korespondensi : *swanjayadaniel@gmail.com*

BAB 1
PENDAHULUAN

1.1 Latar Belakang

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. Diana P, (2011) menggunakan *Hypergraph Partitioning* untuk identifikasi topik. 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.

Ida Ayu, dkk, (2011) menggunakan *Probabilistic Latent Semantic* Analysis (PLSA) untuk mengelompokan kata-kata ke dalam topik-topik yang belum diketahui (*latent*), kemudian menggunakannya untuk mengklasterkan dokumen. Ashari N (2012) menggunakan Latent Semantic Analysis (LSA) dan Singular Value Decomposition (SVD) untuk mengekstraksi topik-topik utama harian dari kumpulan dokumen berita online berbahasa Indonesia.

Latent Dirichlet Allocation (LDA) merupakan teknik pemodelan topik yang berdasarkan *bag of words*. LDA ini melakukan suatu inferensi terhadap dokumen dengan berdasarkan fungsi polinomial. Beberapa teknik *generative model* telah dikembangkan untuk melakukan analisis berdasarkan LDA ini. Yang paling umum digunakan adalah gibbs sampling (Han Xiao, 2010). teknik ini melakukan estimasi topik dan meningkatkan kualitas estimasi hingga terbentuk estimasi topik yang stabil. Indra L, dkk (2014) mengusulkan metode untuk pengurutan kalimat berdasarkan topik kata kunci menggunakan LDA. Metode yang diusulkan memiliki tiga tahapan. Pertama, mengelompokan 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 *Graph* dimana *node*-nya Penulis dan *edge*-nya adanya karya tulis yang pernah ditulis bersama, *Graph* ini disebut *Co-Authorship Graph*. Pada tahun 2008, Vivit WR membuat *Co-Authorship Graph* yang disebut *graph* 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.

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.*

Pengelompokan 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

*Cluster*ing 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.

Bento, C., (2013) mencari komunitas penulis dan hubungannya berdasarkan *co-authorship network*, menggunakan dua dataset, CiNii, informasi biblio*graph*i tentang publikasi ilmiah Jepang sejak tahun 1886, dan DBLP, informasi biblio*graph*i publikasi ilmiah di bidang Ilmu Komputer tahun 1986 sampai dengan 2012. Informasi yang ada pada masing-masing biblio*graph*i 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 Blondel, dkk (2008) 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.

Pengelompokan Penulis yang ada sampai saat ini hanya berdasarkan kolaborasi Penulis, kesamaan atribut karya yang dibuat (*jaccard*), hasil pengelompokan 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 dikelompokan 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 mengelompokan 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 mengelompokan Penulis berdasarkan topik penelitian. Usulan ini dibagi menjadi tiga tahap, ektraksi topik penelitian, pembentukan *Co-Authorship Graph* dan pengelompokan 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 pengelompokan 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*.

1.2 Perumusan Masalah

Perumusan masalah dalam tesis ini adalah sebagai berikut:

1. Bagaimana mengekstraksi topik karya tulis menggunakan *Latent Dirichlet Allocation* (LDA)?
2. Bagaimana membangun *Co-Authorship Graph* berdasarkan kolaborasi dan kesamaan/kemiripan topik penelitian antar Penulis?
3. Bagaimana mengelompokan Penulis menggunakan *Hypergraph Partitioning*?

1.3 Batasan Masalah

Dalam tesis ini, batasan masalah yang dibahas diuraikan sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah data Karya Tulis Ilmiah dari Lembaga Penelitian dan Pengabdian Masyarakat (LPPM) ITS Surabaya.
2. Penulis adalah dosen yang membimbing Tugas Akhir, Tesis ataupun Disertasi.

1.4 Tujuan dan Manfaat Penelitian

Tujuan dari penelitian ini adalah mengelompokan Penulis yang memiliki kesamaan/kemiripan topik karya tulis menggunakan *Hypergraph Partitioning*.

Manfaat dari penelitian ini adalah untuk mengetahui kelompok penelitian atau grup riset yang memiliki kesamaan/kemiripan topik penelitian, sehingga memungkinkan adanya kolaborasi baru antar bidang atau Jurusan dan meningkatkan produktifitas Penulis. Selain itu juga dapat memberikan informasi kepada komunitas penulis pada umumnya untuk bisa berkolaborasi membuat karya tulis bersama-sama atau membentuk kerjasama antar komunitas penulis.

1.5 Kontribusi Penelitian

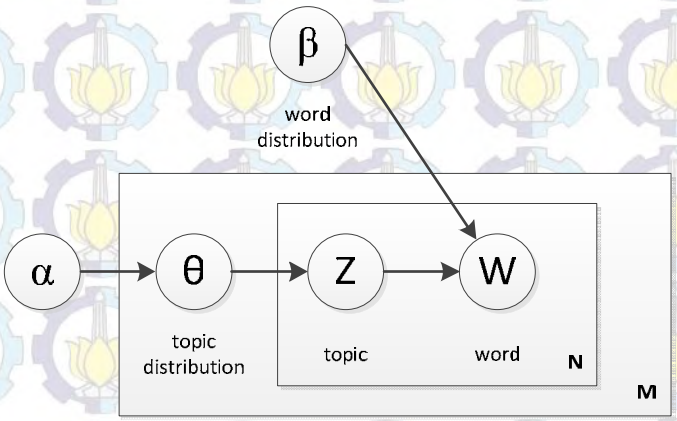
Kontribusi penelitian ini adalah menggunakan *Hypergraph Partitioning* pada *Co-Authorship Graph* untuk mengelompokan Penulis berdasarkan kesamaan/kemiripan topik penelitian.

BAB 2
KAJIAN PUSTAKA

Pada subbab ini, dijelaskan teori yang menjadi dasar pengerjaan tesis. Dasar teori yang diuraikan meliputi *Latent Dirichlet Allocaton* dan *Hypergraph Partitioning*.

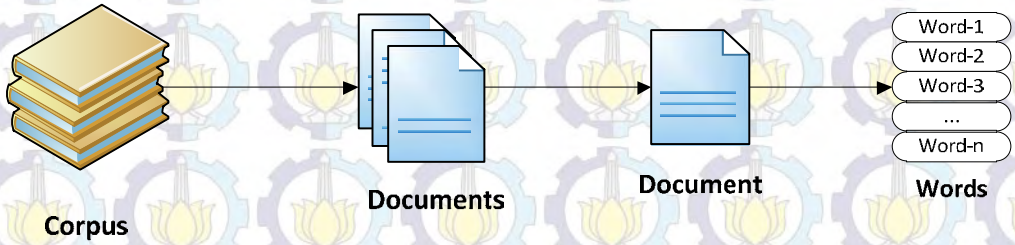
2.1 *Latent Dirichlet Allocaton*

Topik dapat ditentukan dengan memperhatikan beberapa kata-kata penting yang ada pada dokumen dan tidak terletak secara berurutan, kumpulan kata yang tidak berurutan ini disebut *bag-of-words*. Usaha untuk menentukan topik pada dokumen disebut ekstraksi topik, 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.



Gambar 2.1 Pemodelan LDA

Gambar 2.1 menggambarkan model yang mendeskripsikan model generatif ini. Level pertama *corpus-level* parameter, level kedua *document-level* variabel, dan level ketiga *word-level* variabel. Dimana *Word* adalah unit terkecil dari suatu kosakata V. *Document* adalah serangkaian N *words*. *Corpus* adalah himpunan M *documents*. Dengan asumsi *words* pada *document* dapat dipertukarkan dan *documents* juga dapat demikian. Gambar 2.2 menggambarkan hubungan antara *Corpus*, *Documents*, *Document* dan *Words*.



Gambar 2.2 Hubungan *Corpus*, *Documents*, *Document* dan *Words*.

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 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.

Algoritma *Latent Dirichlet Allocation* (LDA) :

1. Jumlah Dokumen = M.

$$Documents = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ \dots \\ d_n \end{bmatrix}$$

2. Vector Vocab, Ekstrak *distinct* kata pada semua dokumen.

$$Vocab = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \dots \\ w_n \end{bmatrix}$$

4. Matriks Word per Dokumen, tidak diunikkan.

$$Word = \begin{Bmatrix} w_{d_1,1} & w_{d_1,2} & \dots & w_{d_1,n} \\ w_{d_2,1} & w_{d_2,2} & \dots & w_{d_1,n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ w_{d_m,1} & w_{d_2,2} & \dots & w_{d_m,n} \end{Bmatrix}$$

dimana :
d_m : Dokumen ke-m
n : Nomor urut *term* pada dokumen
$w_{d_m,n}$: *Term* ke-n pada dokumen ke-m

5. Tentukan jumlah distribusi Topik = *K*.

$$Topics = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \dots \\ z_n \end{bmatrix}$$

6. Gibbs Sampling.
   1. Untuk setiap dokumen d, secara random beri label tiap kata pada dokumen d dengan salah satu topik *K*.

$$Z = \begin{Bmatrix} z_{w_{d_1,1}} & z_{w_{d_1,2}} & \dots & z_{w_{d_1,n}} \\ z_{w_{d_2,1}} & z_{w_{d_2,2}} & \dots & z_{w_{d_2,n}} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ z_{w_{d_m,1}} & z_{w_{d_2,2}} & \dots & z_{w_{d_m,n}} \end{Bmatrix}$$

dimana :
d_m : Dokumen ke-m
n : Indeks *term* pada dokumen d_m
$w_{d_m,n}$: *Term* ke-n pada dokumen ke-m
$z_{w_{d_m,n}}$: Topik dari *Term* ke-n pada dokumen ke-m

2. Distribusi kata, W_i, pada tiap topik, Z_i, NW

$$NW = \begin{Bmatrix} NW_{w_1,Z_1} & NW_{w_1,Z_2} & \dots & NW_{w_1,Z_k} \\ NW_{w_2,Z_1} & NW_{w_2,Z_2} & \dots & NW_{w_2,Z_k} \\ \vdots & \vdots & & \vdots \\ NW_{w_n,Z_1} & NW_{w_n,Z_2} & \dots & NW_{w_n,Z_k} \end{Bmatrix}$$

dimana :

w_n : *Term* ke-n pada Vocab

Z_k : Topik ke-k

NW_{w_n,Z_k} : Banyak *term* ke-n yang berlabel topik ke-k

3. Distribusi topik, Z_i, pada tiap dokumen, d_i, ND

$$ND = \begin{Bmatrix} ND_{d_1,Z_1} & ND_{d_1,Z_2} & \dots & ND_{d_1,Z_k} \\ ND_{d_2,Z_1} & ND_{d_2,Z_2} & \dots & ND_{d_2,Z_k} \\ \vdots & \vdots & & \vdots \\ ND_{d_m,Z_1} & ND_{d_m,Z_2} & \dots & ND_{d_m,Z_k} \end{Bmatrix}$$

dimana :

d_m : Dokumen ke-m

Z_k : Topik ke-k

ND_{d_m,Z_k} : Banyak label topik ke-k pada Dokumen ke-m

4. Jumlah distribusi NW dan ND

$$NWSum_{Z_i} = \sum_{j=1}^{m} NW_{w_j,Z_i}$$

$$NDSum_{d_i} = \sum_{j=1}^{k} ND_{d_i,Z_j}$$

5. Untuk setiap kata, $w_{d_i,j}$, pada Matriks Word per Dokumen

    1. Topik = $Z_{w_{d_i,j}}$
    2. $NW_{w_{d_i,j},topik} = NW_{w_{d_i,j},topik} - 1$
    3. $ND_{d_i,topik} = ND_{d_i,topik} - 1$
    4. $NWSum_{topik} = NW_{Sum_{topik}} - 1$
    5. $NDSum_{d_i} = NDSum_{d_i} - 1$
    6. *Multinomial sampling* menggunakan metode komulatif,

$$s_x = \frac{NW_{w_{d_i,j},x} + \beta}{NWSum_x + n \times \beta} \times \frac{ND_{d_i,x} + \alpha}{NDSum_{d_i} + K \times \alpha}$$

$$P_x = \begin{cases} s_x \; ; x = 1 \\ \sum_{x=2}^{k} s_{x-1} \; ; x > 1 \end{cases}$$

dimana :

d_i	: Dokumen ke-i
j	: Indeks *term* pada dokumen d_i
x	: Indeks sampling topik
β	: Parameter multinomial
α	: Parameter dirichlet
K	: Banyak topik
n	: Banyak *term* pada Vocab
$w_{d_i,j}$	: *Term* ke-j pada dokumen d_i
$NW_{w_{d_i,j},x}$	: Banyak *term* ke-n yang berlabel x
$ND_{d_i,x}$	: Banyak label x pada dokumen d_i
$NWSum_x$	: Jumlah distribusi *term* yang berlabel x
s_x	: Mulitinomial sampling topik ke-x
P_x	: Komulatif Mulitinomial sampling topik ke-x

7. Skala sample, $u = Random \times P_{k-1}$
8. Topik baru untuk kata, $w_{d_i,j}$, adalah indeks terkecil dari P_x yang lebih kecil dari u.
9. $NW_{w_{d_i,j},topikBaru} = NW_{w_{d_i,j},topikBaru} + 1$
10. $ND_{d_i,topikBaru} = ND_{d_i,topikBaru} + 1$
11. $NWSum_{topikBaru} = NW_{Sum_{topikBaru}} + 1$
12. $NDSum_{d_i} = NDSum_{d_i} + 1$

6. Update Parameter,

$$\theta_{d_i,j} = \frac{ND_{d_i,x} + \alpha}{NDSum_{d_i} + K \times \alpha}$$

$$\Phi_{d_i,j} = \frac{NW_{w_{d_i,j},x} + \beta}{NWSum_x + n \times \beta}$$

7. Hasilnya adalah :
   1. matriks Phi, Φ, probabilitas *term* pada Vocab terhadap Topik, dan
   2. matriks Theta, θ, probabilitas Topik terhadap Dokumen.

2.2 Probabilitas Topik

Topik secara umum diartikan sebagai pokok pembicaraan dan berupa kalimat lengkap. Pada LDA topik direpresentasikan sebagai sebuah vector yang berisi probabilitas kemunculan dari semua *term* atau kata yang muncul pada semua Dokumen, Kamus Kata atau Vocabulary, dari probabilitas tersebut didapatkan *term-term* yang dominan, beberapa *term* yang memiliki probabilitas tertinggi. *Term* dominan ini sering digunakan untuk menentukan label dari topik tersebut, tetapi *term-term* tersebut tidak dapat dirangkai menjadi kalimat. Misalnya dari Gambar 2.3 terdapat sebuah Dokumen yang akan ingin diketahui topiknya, setelah diektrak menggunakan LDA didapatkan probabilitas setiap *term* pada Dokumen tersebut, dan 9 *term* yang memiliki probabilitas tertinggi pada tiap topik ditunjukkan pada Tabel 2.1. Berdasarkan kata dominan tersebut dapat disimpulkan bahwa Topik 1 adalah tentang Pembangunan, Topik 2 tentang Asing, Topik 3 tentang Hiburan, dan Topik 4 tentang Indonesia. Tetapi tidak ada aturan baku dalam penentuan label dari tiap topik. Berdasarkan frekuensi tiap topik pada sebuah Dokumen didapatkan probabilitas topik dari Dokumen tersebut, dan topik yang memiliki probabilitas tertinggi dianggap sebagai *main topic* dari Dokumen tersebut.

Perkembangan teknologi dan pembangunan Jepang *term*asuk dalam bidang kebudayaan setidaknya telah memicu peningkatan jumlah pelajar asing yang mempelajari bahasa Jepang *term*asuk dari Indonesia. Hal serupa juga diraih Korea Selatan yang berhasil menajamkan pengaruhnya tidak hanya di bidang teknologi tetapi juga di bidang hiburan. Riak gelombang ekspor kebudayaan Korea Selatan sudah terasa hingga ke seluruh dunia bahkan membanjiri Indonesia mulai dari musik, film, dan bahkan gaya hidup selain tentunya juga bahasanya. Inilah momen yang harus direbut Indonesia dalam menancapkan identitasnya di dalam kompetisi merebut pengaruh masyarakat internasional lewat bahasa; pembangunan secara utuh dan menyeluruh tanpa meninggalkan identitas bangsa. Ini tentunya tidak terlepas dari kualitas sumber daya manusia Indonesia dalam mengembangkan pembangunan di berbagai sektor tanpa menyingkirkan arti penting dari budaya dan bahasa.

Gambar 2.3 Contoh sebuah Dokumen, warna menunjukkan keanggotaan topik.

Misal pada sebuah Dokumen seperti pada Gambar 2.3 setelah diekstrak topiknya didapat distribusi topik seperti pada Tabel 2.1, dimana warna pada Gambar 2.3 menunjukkan keanggotaan kata pada topik.

Tabel 2.1 Topik Dokumen sebelum pelabelan

Label 1	Label 2	Label 3	Label 4
Perkembangan	Jepang	Hiburan	Indonesia
Teknologi	Kebudayaan	Membanjiri	Dunia
Pembangunan	Asing	Musik	Bahasa
bidang	Korea	Film	Momen
peningkatan	Selatan	Gaya	Identitasnya
sumber	Menajamkan	Hidup	Masyarakat
Daya	Pengaruhnya	pelajar	Bangsa
manusia	gelombang	mulai	kualitas
sektor	internasional	pengaruh	budaya

2.3 *Co-Authorship Graph*

Co-Authorship Graph adalah sebuah graph yang merepresentasikan hubungan antar Penulis, dimana node sebagai Penulis dan edge sebagai hubungan antar Penulis. Secara umum *Co-Authorship Graph* digunakan untuk menggambarkan kolaborasi atau kerjasama antar Penulis, apabila terdapat dua Penulis yang pernah menulis bersama, maka pada *Co-Authorship Graph* keduanya akan dihubungkan dan jika tidak pernah bekerjasama, maka keduanya tidak dihubungkan.

Vivit WR (2008) membuat *Co-Authorship Graph* yang disebut *graph* 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.

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*.

2.4 *Hypergraph Partitioning*

Hypergraph adalah suatu bentuk *graph* yang dimana *edge*-nya dapat menghubungkan dua atau lebih verteks yang disebut juga *hyperedges*. Sedangkan *Hypergraph-Partitioning* adalah suatu proses untuk membagi - bagi *hypergraph* ke dalam sub-sub *hypergraph* atau *graph*.

Proses membagi sebuah *graph* G menjadi beberapa himpunan yang saling lepas. Suatu *graph G = {V, E}*, dimana $V = \{ v_1, v_2, \ldots , v_n \}$ yang merupakan node, $E = \{ e_1=\{v_1,v_2\}, e_2=\{v_1,v_3\},\ldots e_n \}$ yang merupakan edge yang menghubungkan antar node, maka himpunan partisinya adalah $P = \{S_1, S_2, S_3, \ldots , S_n\}$ dimana *S* adalah himpunan yang berisi node yang terkelompok pada S, misal $S1 = \{ v_1, v_3, v_5, v_6 \}$.

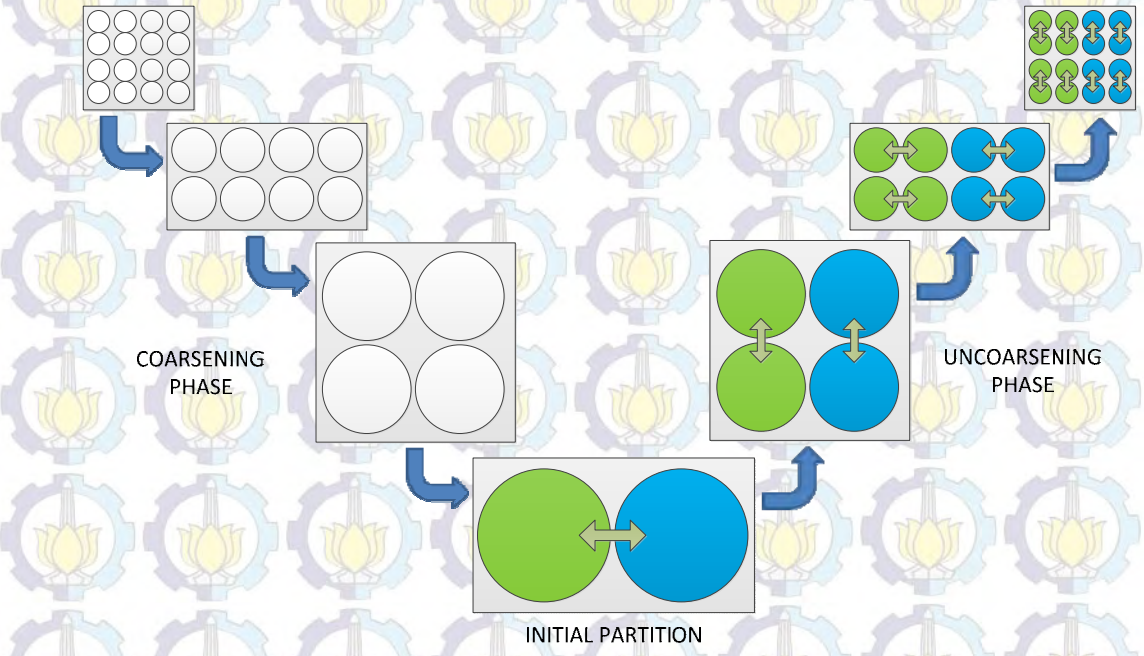
Pada Penelitian David A. Papa (Papa, 2007) proses *Hypergraph-Partitioning* dibagi menjadi tiga proses utama yaitu *coarsening*, *balancing* dan *uncoarsening* (Gambar 2.4). Pertama, fase *coarsening*, *Graph* G disederhakan menjadi dua bagian menggunakan algoritma Maxima Matching, Kedua dilakukan proses *balancing* menggunakan algoritma *Fiduccia-Mattheyeses* sehingga *edge* yang terpotong seminimal mungkin, pada *graph* berbobot algoritma *Fiduccia-Mattheyeses* digunakan untuk mendapatkan memaksimalkan jarak antar partisi atau kelompok. Terakhir fase *uncoarsening*, pada tahap ini node hasil dari fase coarsening yang telah terkelompok dibuka kembali.

2.4.1. *Coarsening*

Pada *coarsening* setahap demi setahap sepasang node yang bertetangga dikelompokan, dimana informasi tiap pengelompokannya disimpan, hingga pada tahap terakhir terbentuk dua partisi atau beberapa apabila antar partisi tidak memiliki hubungan. Kelompok atau node yang menjadi satu kelompok disebut child dari kelompok hasil penggabungan. Tujuan dari *coarsening* adalah menyederhanakan *graph* sehingga terbentuk partisi awal (*Initial Partition*) untuk proses *balancing*.

Algoritma *Maxima Matching (Greedy)* :

1. Untuk tiap node cari tetangga yang memiliki bobot paling tinggi, dimana node tetangga bukan node pasangan dari node lainnya.
2. Kelompokan node yang berpasangan.
3. Hitung bobot antar kelompok hasil proses nomor 2, dimana bobot baru dari kelompok pertama dan kedua adalah bobot terbesar antara node kelompok pertama dan kelompok kedua.
4. Lakukan hingga tidak ada kelompok yang bisa dikelompokan atau terbentuk dua kelompok.



Gambar 2.4 Tiga tahap pada *Hypergraph Partitioning*

2.4.2. ***Balancing***

Tujuan proses *balancing* adalah menyeimbangkan kondisi dua partisi berdasarkan banyaknya edge yang menghubungkan dua partisi dari *graph* tak berbobot atau memaksimalkan jarak antar dua partisi pada *graph* berbobot. Algoritma yang digunakan untuk proses balancing adalah Algoritma *Fiduccia-Mattheyeses*.

Algoritma *Fiduccia-Mattheyeses*, FM (Fiduccia, 1982) 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 dikelompokan 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 :

1. Inisiasi semua node = *unlocked*.
2. Beri label pada kedua partisi, KIRI dan KANAN.
3. Selama ada node = *unlocked*, lakukan,
    a. Hitung nilai gain tiap node pada partisi KIRI, untuk setiap node yang *unlocked*.
    b. Cari node yang memiliki nilai gain terbesar pada partisi pertama, misal LEFT.
    c. Pindahkan LEFT ke partisi KANAN.
    d. LEFT = *locked*.
    e. Hitung nilai gain tiap node pada partisi KANAN, untuk setiap node yang *unlocked*.
    f. Cari node yang memiliki nilai gain terbesar pada partisi KANAN, misal RIGHT.
    g. Pindahkan RIGHT ke partisi KIRI.
    h. RIGHT = *locked*.
    i. Hitung jumlah bobot *edge* yang terpotong, akibat perpindahan LEFT dan RIGHT.

4. Solusi adalah kondisi saat LEFT dan RIGHT yang memiliki jumlah bobot dari *edge* yang terpotong paling besar ditukar.

Dimana :

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

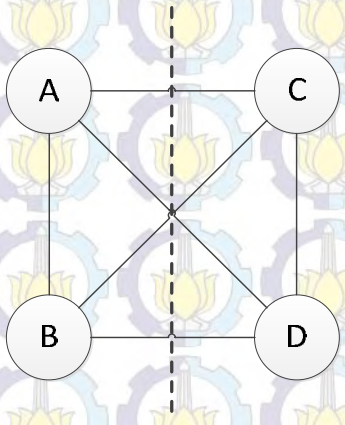
FS(α) = jumlah bobot dari *edge* yang menghubungkan node α dengan node-node yang ada di luar partisi dimana node α berada.

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

Untuk contoh perhitungan gain dari *graph* pada Gambar 2.5, maka gain dari masing-masing nodenya adalah :

Tabel 2.2 Contoh perhitungan gain

Node	**FS**	**TE**	Δg
A	AC + AD = 7	AB = 6	1
B	BC + BD = 3	BA = 6	-3
C	CA + CB = 4	CD = 7	-2
D	DA + DB = 6	DC = 7	-1



Gambar 2.5 Contoh Partisi

Bobot *edge* :

	A	**B**	**C**	**D**
A	0	6	3	4
B		0	1	2
C			0	7
D				0

Berdasarkan perhitungan gain pada Tabel 2.2, maka node yang dipindah adalah node A. Tetapi apabila tidak terdapat nilai gain yang lebih besar dari nol, maka tidak dilakukan perpindahan, karena perpindahan yang dilakukan tidak akan berdampak, hal ini hanya berlaku pada perhitungan gain yang pertama.

2.4.3. *Uncoarsening*

Proses uncoarsening bertujuan untuk mengoptimasi kelompok-kelompok hasil proses coarsening, sehingga setiap pasangan *child* partisi pada tiap level memiliki keseimbangan, dengan melakukan proses *balancing* pada tiap pasangan *child* partisi. Setelah proses *uncoarsening* selesai dilakukan maka berdasarkan parameter banyak maksimal anggota partisi, *max*, dilakukan penyaringan kelompok dari partisi awal hingga akhir sesuai urutan pada proses coarsening, dimana kelompok yang memiliki child, yang beranggotakan kurang dari *max*, kelompok tersebut adalah solusi.

BAB 3
METODE PENELITIAN

3.1 Rancangan Penelitian

Secara umum, penelitian ini diawali dengan studi literatur, pengumpulan data, desain sistem, pengujian sistem, analisis hasil, dan penyusunan laporan. Secara lebih detail, penelitian ini dirancang dengan urutan sebagai berikut:

1. Studi literatur

Dalam studi literatur, dikaji berbagai referensi yang berkaitan dengan ekstraksi topik, *Latent Dirichlet Allocation* (LDA), Teori *Graph*, *Co-Authorship Graph*, *Hypergraph*, analisa hasil *clustering*. Kemudian dicari isu yang sedang dihadapi sekaligus solusi pada kasus yang akan diteliti. Dalam tesis ini isu yang diangkat adalah pengelompokan Penulis berdasarkan topik Dokumen karya tulis berupa Skripsi/Tugas Akhir, Tesis dan Disertasi. Selanjutnya ditentukan metode untuk pemecahan solusi tersebut. Dari metode yang akan digunakan ditemukan kekurangan dan kelebihan dari masing-masing metode.

2. Pengumpulan data

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 :

1. Fakultas Matematika dan Ilmu Pengetahuan (FMIPA)
2. Fakultas Teknologi Industri (FTI)
3. Fakultas Teknik Sipil dan Perencanaan (FTSP)
4. Fakultas Teknologi Kelautan (FTK)
5. Fakultas Teknologi Informasi (FTIF)

Informasi KTI yang digunakan adalah,

1. Judul karya tulis (*Title*)
2. Dosen Pembimbing 1 (*Author_1*)
3. Dosen Pembimbing 2 (*Author_2*)

4. Dosen Pembimbing 3 (*Author_3*)
5. Abstraksi karya tulis (*Abstact*)

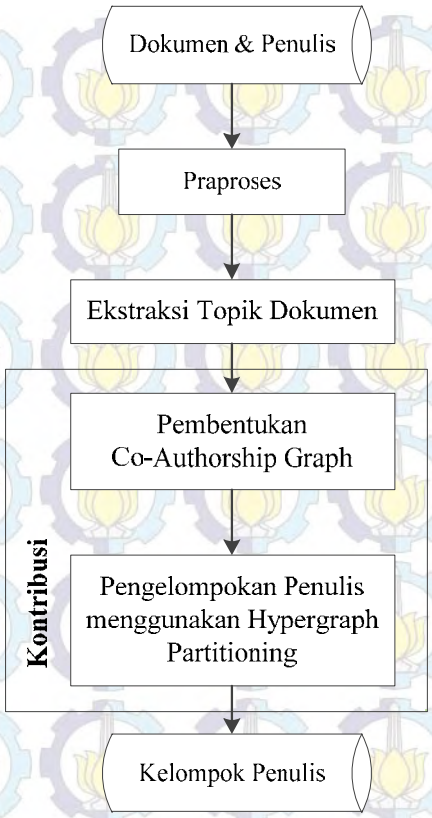
Tabel 3.1 Banyak karya tulis pada tiap Jurusan di ITS Surabaya

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**

Tabel 3.2 Banyak karya tulis berdasarkan banyaknya penulis, tanpa mahasiswa

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

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 (Tabel 3.2).



Gambar 3.1 Arsitektur sistem yang diusulkan

3. Desain sistem

Desain sistem yang diusulkan pada tesis ini terdiri dari empat tahap (Gambar 3.1), yaitu :

1) Praproses

Sebelum data teks diolah, perlu dilakukan pembersihan supaya teks yang ada pada dokumen seragam dan mudah untuk dibaca oleh sistem. Praproses terbagi menjadi 5 tahap, yaitu *lexical analysis*, *Tokienizing*, eliminasi *stopwords*, *stemming* dan pembobotan kata (Gambar 3.3). Dokumen yang diolah adalah Judul dan Abstraksi karya tulis.

a. Lexical analysis

Lexical Analysis adalah proses perubahan karakter-karakter teks yang terdapat pada dokumen ke dalam bentuk kata-kata yang merupakan kandidat *index term*. Tujuan dari *lexical analysis* adalah identifikasi kata-kata dalam teks. Sekilas nampaknya yang dilakukan hanyalah pengenalan spasi yang menjadi pemisah antar kata. Namun prosesnya tidak sesederhana itu,

terdapat empat hal yang perlu diperhatikan yaitu angka, tanda hubung, tanda baca, dan besar kecilnya huruf.

Angka bukanlah index *term* yang baik karena tanpa adanya konteks di sekitarnya maka angka menjadi tidak jelas. Namun terkadang angka memiliki makna yang penting, misalnya nomor kartu kredit yang harus dipertimbangkan sebagai index *term*. Penanganan awal terhadap angka adalah menghilangkan menghilangkan semua kata yang mengandung angka berurutan, kecuali ditentukan aturan lainnya dengan regular expression.

Pertimbangan lainnya adalah tanda hubung. Pemisahan kata-kata yang dihubungkan dengan tanda hubung mungkin berguna untuk mengatasi ketidakkonsistenan penggunaan. Hal ini menyebabkan "state-of-art" sama dengan "state of art". Tetapi ada pula kata-kata dimana tanda hubung memang dibutuhkan, misalnya "glit-edge".

Pada umumnya, tanda baca dihilangkan dalam proses *lexical analysis*. Besar kecilnya huruf, yang merupakan hal keempat yang perlu diperhatikan, juga tidak berpengaruh pada pengidentifikasian *index term*. *Lexical analysis* mengubah semua teks menjadi huruf besar atau huruf kecil.

b. Tokenizing

Tokenisasi merupakan proses pemisahan suatu rangkaian karakter berdasarkan karakter spasi. Token seringkali disebut sebagai istilah (*term*) atau kata, sebagai contoh sebuah token merupakan suatu urutan karakter dari dokumen tertentu yang dikelompokan sebagai unit semantik yang berguna untuk diproses (Salton, 1989).

c. Eliminasi *stopwords*

Kata-kata yang terlalu sering muncul dalam dokumen-dokumen bukanlah pembeda yang baik. Bahkan kata-kata yang muncul 80% dalam dokumen-dokumen tidak berguna dalam proses *text mining*. Kata-kata ini disebut dengan istilah *stopwords* dan umumnya tidak dijadikan *index term*. Kandidat umum stopword adalah *article*, preposisi, dan konjungsi.

Eliminasi *stopwords* menyebabkan pengurangan ukuran strukur index hingga 40%. Karena pengurangan ukuran index, beberapa kata kerja, kata sifat, dan kata keterangan lainnya dapat juga dapat dimasukkan juga ke dalam daftar *stopwords*.

d. *Stemming*

Stemming merupakan sebuah proses yang melakukan *mapping* berbagai variasi morfologikal suatu kata menjadi bentuk dasar kata tersebut. Proses ini disebut juga dengan istilah *conflation*. Berdasarkan pada asumsi bahwa *term-term* yang memiliki bentuk dasar (*stem*) yang sama pada umumnya memiliki makna yang mirip.

Proses *stemming* (Gambar 3.2) dilakukan dengan cara menghilangkan semua imbuhan (*affixes*) baik yang terdiri dari awalan (*prefixes*), sisipan (*infixes*), akhiran (*suffixes*) dan *confixes* (kombinasi dari awalan dan akhiran) pada kata turunan. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi bahasa Indonesia yang benar (Tala, 2003)

Struktur derivasional pada Bahasa Indonesia terdiri atas awalan, akhiran, dan pasangan kombinasi awalan dan akhiran (*konfiks*). Awalan-awalan dalam Bahasa Indonesia yang paling umum digunakan antara lain “ber-”, “di-”, “ke-”, “meng-”, “peng-”, “per-”, dan “ter-”. Awalan-awalan “ber-”, “meng-”, “peng-”, dan “ter-” dapat muncul dalam berbagai bentuk. Bentuk dari setiap awalan ini bergantung kepada karakter pertama dari kata yang mengikuti. Berbeda halnya dengan struktur infleksional, ejaan kata bisa berubah pada saat kata tersebut digabungkan dengan awalannya. Struktur derivasional juga mengenal adanya konfiks, dimana kombinasi dari awalan dan akhiran bergabung dalam satu kata menjadi sebuah kata baru.

Tetapi tidak semua kombinasi awalan dan akhiran yang telah disebutkan dapat disatukan menjadi konfiks. Ada beberapa kombinasi awalan dan akhiran yang tidak valid untuk dikombinasikan menjadi konfiks. Tabel 3.3 menunjukkan daftar kombinasi awalan dan akhiran yang tidak boleh dilakukan.

Selain itu, tata bahasa morfologikal dalam Bahasa Indonesia juga mengenal adanya sisipan seperti “-em-”, “-el-”, “-er-”, dan “-in-”. Contoh penggunaan sisipan misalnya, kata “gemuruh” yang berasal dari kata “guruh” ditambahkan dengan sisipan “-em-”, dan kata “geligi” yang berasal dari kata “gigi” ditambahkan dengan sisipan “-el-”.

Kemungkinan yang terakhir untuk struktur kata yang ada dalam Bahasa Indonesia adalah menambahkan akhiran infleksional untuk kata yang telah memiliki awalan derivasi, kata yang telah memiliki akhiran derivasi, kata yang telah memiliki konfiks derivasi, atau bahkan kata yang telah memiliki awalan ganda. Bentuk ini merupakan bentuk paling kompleks dalam Bahasa Indonesia, yang disebut juga akhiran ganda.

Tabel 3.3 Kombinasi Awalan dan Akhiran yang tidak valid

Awalan	**Akhiran**
ber-	-i
di-	-an
ke-	-i
ke-	-kan
meng-	-an
peng-	-i
peng-	-kan
ter-	-an

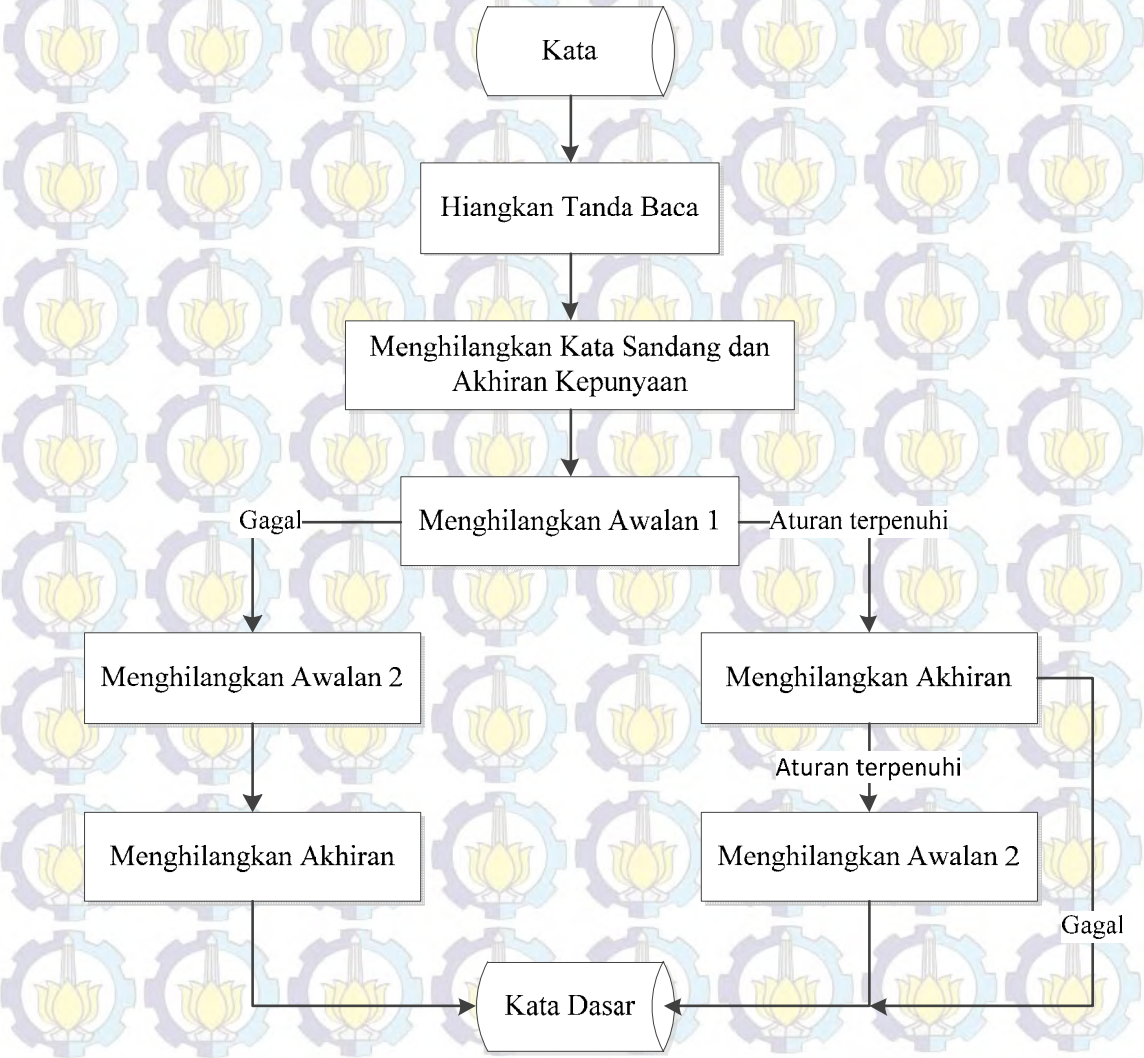
e. Pembobotan Kata menggunakan tf*idf

Tf*Idf adalah perhitungan yang mengGambarkan seberapa pentingnya kata (*term*) dalam sebuah dokumen dan korpus (sekumpulan dokumen). TF atau *Term Frequency* adalah ukuran seringnya kemunculan sebuah *term* dalam sebuah dokumen dan juga dalam seluruh dokumen di dalam korpus, TF dihitung menggunakan persamaan (1). IDF atau Inverse Document Frequency adalah logaritma dari rasio jumlah seluruh dokumen dalam korpus dengan jumlah dokumen yang memiliki *term* yang dimaksud, seperti yang dituliskan secara matematis pada persamaan (2). Bobot kata didapatkan dengan mengalikan kedua persamaan, yang diformulasikan pada persamaan (3), kata yang memiliki bobot kata kecil dieliminasi.

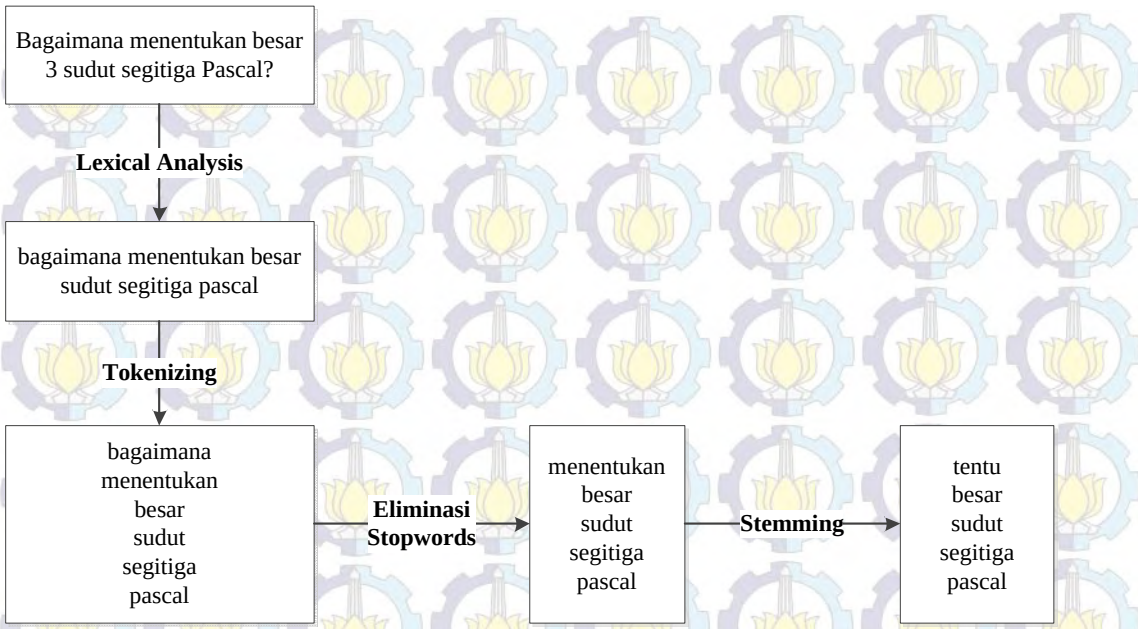
$$tf(i) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)} \quad (1)$$

$$idf_i = \log \frac{|D|}{|\{d: t_i \in d\}|} \quad (2)$$

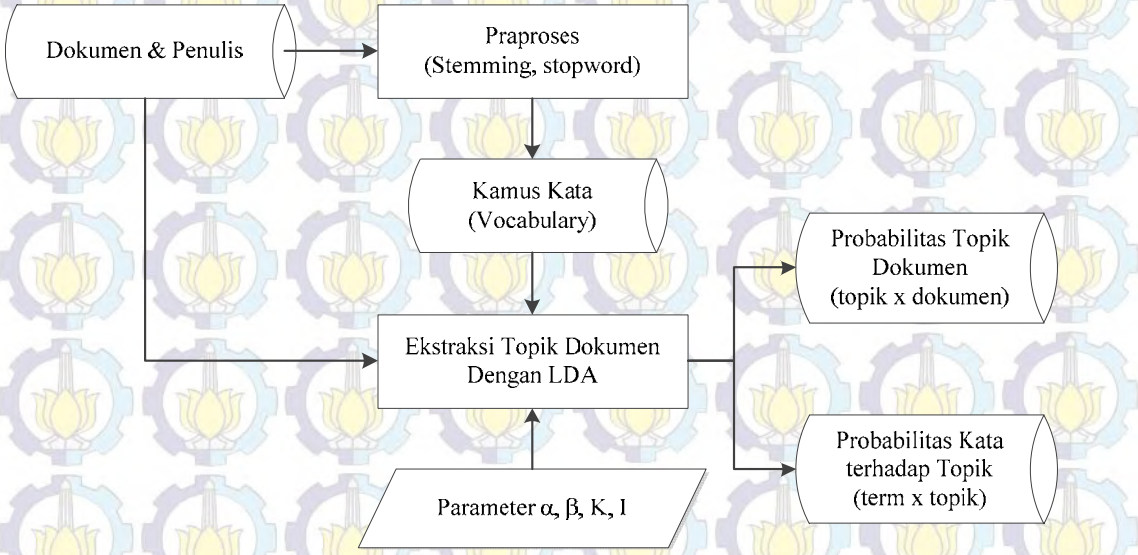
$$tf.idf_{ij} = tf_i(d_j) . idf_i \quad (3)$$



Gambar 3.2 Proses stemming algoritma Tala (Tala, 2003)



Gambar 3.3 Praproses Dokumen



Gambar 3.4 Proses ekstraksi Topik Dokumen

2) Ekstraksi Topik Dokumen

Ekstraksi Topik Dokumen adalah proses mengetahui komposisi probabilitas topik terhadap Dokumen menggunakan Latent Dirichlet Allocation (LDA). Proses Ekstraksi Topik Dokumen (Gambar 3.4) membutuhkan Kamus Kata atau *Vocabulary* (*v*), parameter multinomial (β) dan dirichlet (α), banyak topik (*k*) dan banyak iterasi proses (*I*). Kamus Kata adalah kumpulan kata atau *term* yang

muncul pada Dokumen. Parameter multinomial mempengaruhi distrbusi kata pada setiap topik, parameter dirichlet mempengaruhi disribusi topik pada setiap Dokumen, dan banyaknya iterasi mempengaruhi keakuratan distribusi topik. Hasil dari ekstraksi topik dokumen ini adalah matriks probabilitas *term* terhadap topik atau matriks Phi dan matriks probabilitas topik terhadap Dokumen atau matriks Theta. *Term* yang memiliki probabilitas tertinggi di tiap topik pada matriks Phi, merepresentasikan *term* dominan dari tiap topik (Tabel 3.4). Topik yang memiliki probabilitas tertinggi di satu Dokumen pada matriks Theta, merepresentasikan topik dominan dari Dokumen tersebut (Tabel 3.5).

3) Pembentukan *Co-Authorship Graph*

Co-Authorship Graph digunakan untuk menggambarkan hubungan antar Penulis, dimana Penulis sebagai *node* dan hubungan antar penulis sebagai *edge*. Proses pembentukkan *Co-Authorship Graph* terdiri dari tiga proses, yaitu perhitungan bobot kerjasama atau kolaborasi antar Penulis, perhitungan bobot kesamaan topik antar Penulis, perhitungan bobot edge (Gambar 3.5).

a. Perhitungan bobot kolaborasi

Bobot kolaborasi antara penulis A dan penulis B diperoleh dari rasio banyaknya Dokumen yang ditulis oleh secara bersama oleh Penulis A dan B dengan banyaknya keseluruhan Dokumen yang ditulis oleh Penulis A tanpa Penulis B, Penulis B tanpa Penulis A, dan yang ditulis bersama. Bobot kolaborasi ini dihitung menggunakan formula Jaccard, seperti pada persamaan (4). Apabila Penulis A dan Penulis B selalu berkolaborasi maka nilai Jaccard antara penulis A dan Penulis B adalah 1, tetapi bila Penulis A dan Penulis B tidak pernah berkolaborasi maka nilai Jaccard antara penulis A dan Penulis B adalah nol. Akan tetapi pada penelitian ini bobot kolaborasi antara Penulis juga diperhitungkan sehingga nilainya tidak boleh nol, maka bobot kolaborasi didapat dari nilai Jaccard ditambah dengan suati nilai positif, yaitu 0,1, seperti pada persamaan (5), sehingga apabila terdapat dua Penulis yang tidak pernah berkolaborasi maka bobot kolaborasinya adalah 0,1 dan bila selalu berkolaborasi bobot kolaborasinya adalah 1,1.

Tabel 3.4 Sepuluh kata yang dominan berdasarkan matriks Phi

Topik-0	%	Topik-1	%	Topik-2	%	Topik-3	%
file	5,26	ekstrak	2,59	berita	4,48	*cluster*	7,41
kompres	3,60	dimensi	1,75	ruang	2,73	kelompok	5,08
pesan	3,60	tekstur	1,75	produk	2,62	mean	2,75
aman	3,23	komponen	1,67	distribusi	2,55	batik	2,20
basisdata	2,96	warna	1,46	sumber	2,02	kelas	2,02
enkripsi	2,08	matriks	1,44	rekomendasi	1,88	fuzzy	1,85
kode	1,56	wavelet	1,44	konten	1,76	efektif	1,76
steganog*raph*i	1,49	fuzzy	1,44	daya	1,73	dataset	1,48
Gambar	1,32	kernel	1,42	agen	1,71	genetic	1,36
smartphone	1,30	wajah	1,39	digital	1,71	kategori	1,13

Topik-4	%	Topik-5	%	Topik-6	%	Topik-7	%
fuzzy	2,76	bisnis	5,26	noise	5,26	protokol	5,26
prediksi	2,76	manajemen	3,60	tree	3,60	video	3,60
kanker	2,19	orientasi	3,60	kualitas	3,60	node	3,60
network	2,04	architect	3,23	filter	3,23	suara	3,23
vector	1,89	domain	2,96	jadwal	2,96	route	2,96
adapt	1,81	desain	2,08	atribute	2,08	paket	2,08
latih	1,46	lapor	1,56	struktur	1,56	rute	1,56
machine	1,40	integrasi	1,49	mahasiswa	1,49	voip	1,49
support	1,36	kelola	1,32	decision	1,32	stream	1,32
neural	1,36	transaksi	1,30	bidang	1,30	simulasi	1,30

Topik-8	%	Topik-9	%	Topi-10	%	Topik-11	%
lokasi	3,60	dokumen	8,38	segmen	8,41	main	8,00
posisi	3,18	bahasa	3,57	retina	2,46	gerak	5,10
indonesia	2,59	teks	2,91	langkah	2,16	game	4,10
user	2,49	topik	2,64	threshold	2,04	robot	2,81
google	2,03	spesifik	1,80	buluh	2,04	real	2,27
social	1,98	semantik	1,55	transform	2,01	sensor	1,66
peta	1,85	ajar	1,54	titik	1,70	simulasi	1,27
situs	1,79	bobot	1,49	gigi	1,59	lingkungan	1,07
wisata	1,52	konsep	1,40	darah	1,56	program	1,04
foto	1,51	kalimat	1,40	fundus	1,45	kendara	1,00

Tabel 3.5 Matriks Probabilitas topik pada tiap dokumen (20 Dokumen pertama dari Jurusan Teknik Informatika, 12 Topik)

NRP	T-00	T-01	T-02	T-03	T-04	T-05
5101109048	2,66	4,8	3,39	5,04	39,9	3,33
5102109036	3,25	8,94	34,88	9,12	5,92	6,86
5102109046	3,65	15,13	6,55	22,58	3,25	5,47
5103100004	5,45	3,62	3,4	4,8	3,63	44,15
5103100029	4,03	8,44	5,07	51,1	3,85	3,91
5103100063	7,45	21,51	7,29	10,13	8,74	4,35
5103100082	3,62	6,23	3,64	48,71	3,55	4,87
5103100087	4,33	8,72	2,91	22,96	21,61	9,27
5103109040	55,32	4,36	3,75	4,12	3,37	3,42
5104100008	6,43	3,98	3,2	4,31	5,02	3,46
5104100010	5,56	4,38	29,44	3,57	4,39	8,63
5104100026	21,51	2,44	43,82	3,07	2,72	3,08
5104100028	4,04	4,85	5,91	5,06	6,6	4,08
5104100030	11,49	4,49	27,79	5,43	3,99	10,62
5104100033	3,07	56,85	5,21	4,68	4,1	3,43
5104100039	6,24	9,1	6,72	6,61	4,34	6,59
5104100044	35,85	3,72	10,85	3,76	3,12	3,15
5104100048	56,37	2,79	4,15	3,08	2,95	7,17
5104100049	9,17	3,55	8,64	5,83	3,96	3,69
5104100051	4,15	6,09	5,79	3,85	4,89	3,73
5104100052	6,28	4,16	25,81	19,86	3,14	10,31
5104100062	3,67	10,55	2,67	3,59	4,18	3,96
5104100068	3,23	5,22	12,15	28,23	12,7	4,74
5104100078	5,97	4,53	5,97	5,23	4,3	7,42
5104100085	2,96	7,76	2,82	4,88	7,51	3,39

NRP	T-06	T-07	T-08	T-09	T-10	T-11
5101109048	2,89	5,98	23,22	2,66	2,4	3,73
5102109036	3,66	5,91	8,04	5,24	3,29	4,89
5102109046	4,49	14,65	4,56	2,93	9,79	6,95
5103100004	8,87	8,22	2,77	2,57	4,89	7,64
5103100029	4,16	3,66	5,64	3,39	3,46	3,32
5103100063	8,31	3,98	6,21	4,62	10,63	6,78
5103100082	6,45	4,01	4,77	3,5	3,95	6,69
5103100087	4,77	6,61	4,24	2,99	7,39	4,21

NRP	T-06	T-07	T-08	T-09	T-10	T-11
5103109040	3,24	7,46	4,14	3,66	3,52	3,65
5104100008	3,65	50,31	7,2	3,33	2,44	6,66
5104100010	3,46	3,28	10,59	5,07	2,81	18,81
5104100026	3,66	7,8	2,5	3,2	2,42	3,79
5104100028	47,63	4,1	4,18	4,12	5,49	3,95
5104100030	8,14	4,83	8,48	3,82	4,24	6,68
5104100033	3,81	3,2	3,53	4,1	3,64	4,37
5104100039	10,07	7,37	4,13	4,13	5,14	29,56
5104100044	5,05	16,22	7,08	4,28	3,12	3,81
5104100048	3,9	3,44	5,42	4,06	2,59	4,06
5104100049	3,87	35,99	6,21	5,2	3,48	10,42
5104100051	3,92	18,15	6,7	3,77	3,45	35,5
5104100052	10,93	3,44	5,32	3,47	3,5	3,76
5104100062	3,4	2,86	2,75	2,93	2,75	56,67
5104100068	9,53	3,47	6,48	5,17	5,38	3,71
5104100078	11,45	12,71	20,55	7,44	3,64	10,79
5104100085	3,16	5,4	2,9	11,72	2,94	44,56

$$Jaccard(A,B) = \frac{Dokumen\ (A) \cap Dokumen\ (B)}{Dokumen\ (A) \cup Dokumen\ (B)} \quad (4)$$

$$Kolaborasi(A,B) = Jaccard(A,B) + 0{,}1 \quad (5)$$

$$Similarity(A,B) = \frac{A.B}{||A||||B||} \quad (6)$$

$$Bobot(A,B) = c \times Similarity\ (A,B) + (1-c) \times Kolaborasi(A,B) \quad (7)$$

dimana,

$Dokumen\ (A)$: Dokumen yang ditulis oleh A.

$Dokumen\ (B)$: Dokumen yang ditulis oleh B.

c : Konstanta

b. Perhitungan bobot kesamaan topik Penulis

Bobot kesamaan topik Penulis didapatkan dari nilai Cosine Similarity dari probabilitas topik Penulis, dimana probabilitas topik Penulis adalah representasi dari jumlah probabilitas tiap topik dari Dokumen yang

ditulis oleh Penulis tersebut (Tabel 3.6). Bobot kesamaan topik antara Penulis A dan Penulis B dihitung dengan persamaan (6), dimana jika topik kedua Penulis sangat mirip akan menghasilkan nilai 1 dan 0 jika berbeda sama sekali.

c. Perhitungan bobot edge

Pada *Co-Authorship Graph*, edge merepresentasikan hubungan antar peneliti, dan pada penelitian ini hubungan tersebut memiliki bobot yang diperoleh dari bobot kolaborasi dan bobot kesamaan topik. Bobot edge yang menghubungkan antara Penulis A dan Penulis B, dihitung dengan persamaan (7). Untuk mempermudah perhitungan maka bobot edge hanya akan memiliki rentang nilai antara 0 sampai 1, sehingga rentang nilai bobot kolaborasi dan bobot kesamaan topik juga harus dibuat menjadi 0 sampai 1, dengan cara membagi nilai bobot dengan nilai maksimal dari keseluruhan bobot.

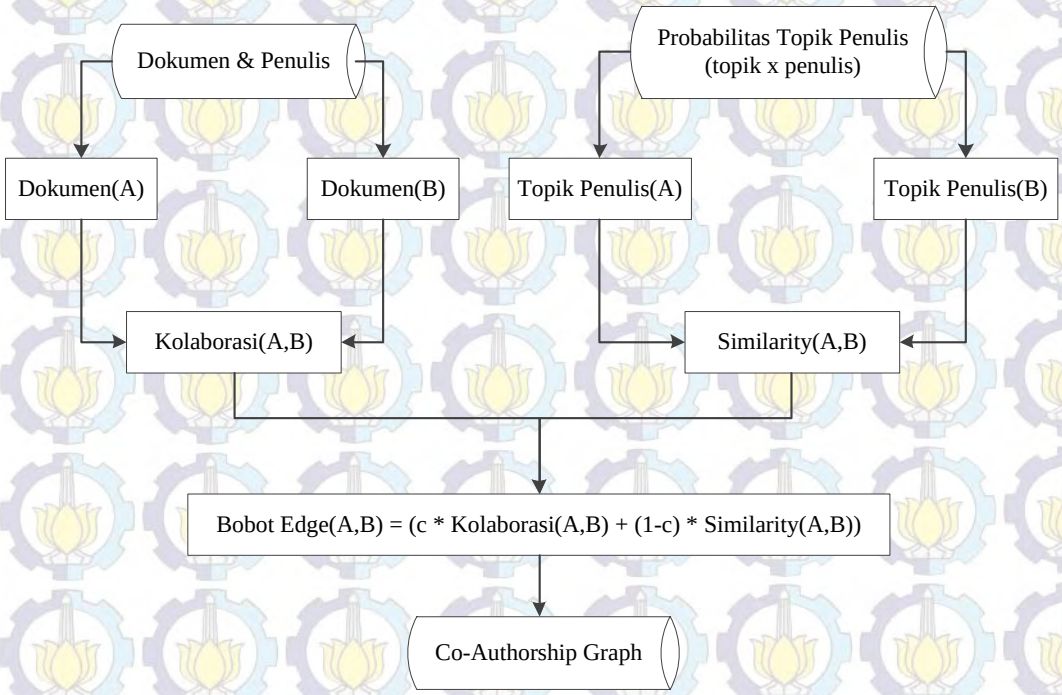
Tabel 3.6 Perhitungan Probabilitas Topik Penulis

ID Dokumen	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
5105100076	4,52	4,48	3,45	3,56	3,27	25,44	2,31	21,47	7,33	6,19	15,51	2,47
5105201004	49,62	3,69	6,58	3,85	3,63	7,18	3,03	3,12	3,22	3,00	3,01	10,10
5106100022	34,03	2,18	4,14	2,87	2,44	40,27	2,13	2,31	2,25	2,21	2,32	2,86
5106100040	2,86	4,31	2,20	2,76	7,97	3,56	2,19	3,75	62,59	2,26	2,42	3,12
5106100043	11,09	3,38	2,64	2,59	7,88	2,53	2,49	2,79	52,89	6,74	2,44	2,55
5106100048	36,28	4,47	4,54	4,89	4,54	17,88	4,64	4,49	4,49	4,54	4,69	4,52
5106100059	16,42	8,24	35,54	4,97	2,80	3,86	6,03	2,77	3,18	4,43	9,10	2,67
5106100075	2,95	2,73	4,89	2,95	2,99	54,63	2,76	2,66	15,09	2,68	2,86	2,82
5106100090	3,48	3,14	2,89	3,91	5,57	2,89	2,77	3,26	63,40	2,97	2,91	2,80
5106100095	44,15	5,06	3,56	9,32	4,98	11,22	2,97	3,04	2,88	4,58	4,09	4,15
Jumlah	205,4	41,68	70,43	41,67	46,07	169,46	31,32	49,66	217,32	39,6	49,35	38,06
Normalisasi	**20,54**	**4,17**	**7,04**	**4,17**	**4,61**	**16,95**	**3,13**	**4,97**	**21,73**	**3,96**	**4,93**	**3,81**

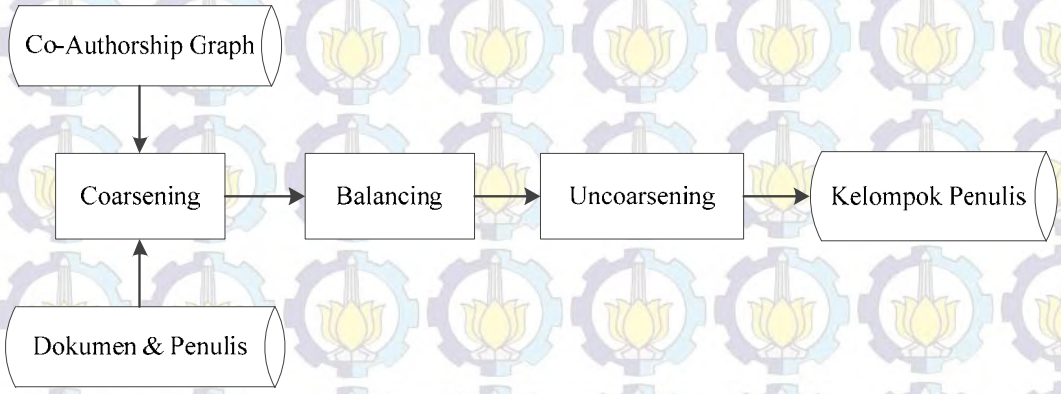
4) Pengelompokan Penulis menggunakan *Hypergraph-Partitioning*

Proses pengelompokan Penulis terdiri dari 3 tahap (Gambar 3.6). Coarsening, inisiasi partisi dimana semua node dikelompokan menjadi 2 kelompok atau lebih jika pada saat pengelompokan terdapat beberapa kelompok yang tidak memiliki hubungan. *Balancing* menggunakan algoritma Fiduccia-Mattheyeses,

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.



Gambar 3.5 Pembentukan *Co-Authorship Graph*.



Gambar 3.6 Pengelompokan Penulis menggunakan *Hypergraph-Partitioning*

3.2 Rancangan Pengujian

Untuk menguji apakah metode yang diajukan bisa berjalan dengan baik perlu dilakukan beberapa kali ujicoba. Ada beberapa kombinasi skenario pengujian yang akan dilakukan.

1. Pengaruh banyak topik dan *c* terhadap kualitas kelompok.

   Pengujian ini bertujuan untuk mengetahui nilai *K* (banyak topik) dan *c* (nilai konstanta bobot) yang menghasilkan pengelompokan yang terbaik. Pada pengujian ini dilakukan beberapa skenario :

   a. Penulis dari satu Jurusan

   b. Penulis dari beberapa Jurusan

   Hasil dilihat dari nilai *Average Silhouette Width* (ASW) masing masing pengelompokan.

2. Pengaruh banyak topik dan *c* terhadap validitas kelompok.

   Pengujian ini bertujuan untuk mengetahui nilai *K* (banyak topik) dan *c* (nilai konstanta bobot) yang menghasilkan pengelompokan yang terbaik untuk Penulis di lingkungan ITS, Sebagai data pembanding digunakan data LPPM tentang Penelitian tahun 2012 dan 2013, anggota Laboratorium, Hibah Penelitian dan LBE (Lab based Education). Hasil dilihat dari nilai *Entropy* masing-masing pengelompokan.

3.3 Metode Pengujian

Hasil uji coba akan dievaluasi sehingga dapat dilihat kinerja metode yang diajukan. Ukuran evaluasi yang digunakan adalah *Entropy* dan *Silhouette* Coefficient. *Entropy* adalah suatu parameter yang menunjukkan tingkat kemurnian dari klaster yang terbentuk. *Entropy* (Zhao, 2001) dihitung berdasarkan matriks confusion hasil klasterisasi dalam persamaan :

$$E(C_i) = -\frac{1}{\log q}\sum_{j=1}^{q}\frac{n_i^j}{n_i}\log\frac{n_i^j}{n_i} \quad (8)$$

$$Entropy = \sum_{i=1}^{N}\frac{n_i}{n}E(C_i) \quad (9)$$

Dimana :

$E(C_i)$: Nilai *Entropy Cluster* ke-i

q : Banyak class pada GroudTruth.

n_i^j : Banyak anggota *cluster* ke-i yang *term*asuk class ke-j

n_i : Banyak anggota *cluster* ke-i

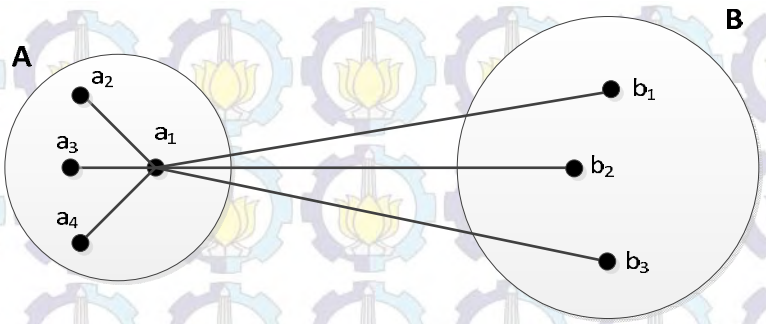
n : Banyak seluruh anggota.

N : Banyak *cluster*.

Semakin besar nilai *Entropy* mengindikasikan nilai validitas kelompok yang semakin baik.

Silhouette Coefficient dikembangkan pertama kali oleh Kaufman dan Rousseeuw (Rousseeuw, 1987). *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* dari sebuah objek Oj berada pada rentang antara -1 sampai dengan 1. Semakin dekat nilai *Silhouette* objek Oj ke 1, maka semakin tinggi derajat Oj di dalam klaster. Berdasarkan penggunaannya, berikut akan dijelaskan cara untuk menentukan *Silhouette coefficient* dari klaster :



Gambar 3.7 Ilustrasi *Silhouette*

Pada Gambar 3.7, terdapat sebuah titik x (a_1) di klaster A, lalu a(x) adalah jarak rata-rata antara titik a_1 dengan titik lain di klaster A, dan b(x) adalah nilai terbesar dari jarak rata-rata antara titik a_1 dengan titik-titik di klaster lainnya. Kemudian hitung nilai *Silhouette* setiap objek dengan persamaan :

$$b(x) = \max_{C_i \neq A} d(x, C_i) \tag{10}$$

$$S_i = \frac{a(x) - b(x)}{max\big(a(x), b(x)\big)} \tag{11}$$

$$ASW = \frac{1}{N}\sum_{i=1}^{N} S_i \tag{12}$$

Nilai *Silhouette* (S_i) akan mengindikasikan derajat kepemilikan tiap objek sebagai berikut :

Tabel 3.7 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.

Setelah mendapatkan nilai *Silhouette* tiap objek dalam klaster, kita dapat menentukan nilai rata-rata *Silhouette cluster* dengan menghitung rata-rata nilai *Silhouette* semua objek yang berada dalam klaster, sehingga diketahui struktur internal *cluster*, sebaik apa kedekatan atau persamaan yang ada dalam kelompok tersebut dan bagaimana keterpisahan antar kelompok. *Average Silhouette width* (ASW) adalah nilai rata-rata dari *Silhouette* dari semua obyek, berdasarkan eksperimen yang dilakukan oleh Rousseeuw (Rousseeuw, 1987), beliau mengemukakan interpretasi terhadap nilai ASW seperti pada Tabel 3.8.

Dari nilai *Silhouette* obyek, dihitung nilai rata-rata *Silhouette width* (ASW) yang mengindikasikan kualitas klastering. Berdasarkan nilai ASW dapat diketahui sebaik apa struktur pengelompokan yang terbentuk. Seperti pada Tabel 3.7, jika ASW berada pada *range* 0,71 sampai 1, maka diketahui bahwa struktur pengelompokan yang terbentuk sangat baik dimana perbedaan antar kelompok sangat baik dan persamaan dalam kelompok sangat tinggi. Jika ASW bernilai lebih dari atau sama dengan 0,51 dan kurang dari 0,71, maka diketahui bahwa struktur pengelompokan sudah baik. Jika ASW bernilai lebih atau sama dengan dari 0,26 dan kurang dari 0,51, maka struktur pengelompokan sudah cukup baik, tetapi jika nilai ASW kurang dari 0,26 maka struktur pengelompokan kurang baik, dimana nilai perbadingan rata-rata *intra-cluster* dan rata-rata *inter-cluster* kecil.

Tabel 3.8 Inteprestasi nilai *Average Silhouette Width* (ASW)

ASW	**Intepretasi**
$0.71 \leq ASW \leq 1$	*A strong structure has been found*
$0.51 \leq ASW < 0.71$	*Reasonable A reasonable structure has been found*
$0.26 \leq ASW < 0.51$	*The structure is weak and could be artificial. Try additional methods of data analysis.*
$ASW < 0.26$	*No substantial structure has been found*

BAB 4
IMPLEMENTASI DAN PENGUJIAN

Pada sub bab ini akan membahas tentang implementasi, pengujian dan pembahasan terkait penelitian yang diusulkan. Tahapan implementasi yang dilakukan sesuai dengan alur pada Gambar 3.1 yang terdiri dari ekstraksi topik penelitian dengan LDA, pembentukkan co-aouthorship *graph*, dan pengelompokan Penulis menggunakan *Hypergraph Partition*. Tahap berikutnya adalah pengujian dari implementasi yang telah dilakukan. Skenario pengujian sesuai dengan skenario yang telah direncanakan sebelumnya pada sub bab 3.4 tentang perancangan pengujian. Tahapan tekahir dari bab ini adalah pembahasan tentang hasil dan evaluasi *Hypergraph Partition* untuk pengelompokan penulis.

4.1 Perangkat Pengujian

Untuk melakukan implementasi dan pengujian *Hypergraph Partition* untuk pengelompokan penulis, penulis menggunakan beberapa perangkat yang terdiri dari perangkat keras dan perangkat lunak.

1. Perangkat Keras

   Perangkat keras yang digunakan untuk implementasi dan pengujian adalah satu buah laptop dengan spesifikasi processor Intel Core i5-2410M 2.3 GHz, RAM 8 Gb.

2. Perangkat Lunak

   Perangkat lunak yang digunakan pada tahapan implementasi dan pengujian adalah Sistem operasi Windows 7 64 bit dan aplikasi NetBeans 8.0.1 dengan JDK versi 8.

4.2 Implementasi Sistem

4.2.1 Dataset

Dataset yang digunakan adalah Karya Tulis mahasiswa ITS berupa Skripsi, Tesis dan Disertasi dari tahun 2005 sampai dengan 2012. Informasi yang digunakan adalah NRP Mahasiswa, Dosen Pembimbing, Judul Karya Tulis dan Abstraksi. Untuk selanjutnya Judul Karya Tulis dan Abstraksi akan disebut sebagai Dokumen,

Dosen Pembimbing disebut sebagai Penulis, dan NRP Mahasiswa akan disebut sebagai ID Dokumen. Tabel 4.2 menunjukkan representasi Data Dokumen yang digunakan pada penelitian ini.

Pada tahap praproses semua Dokumen dibersihkan terlebih dahulu, dengan cara :

a. *Lexical Analysis*, pada tahap ini semua kata yang mengandung angka, semua tanda hubung, tanda baca dihilangkan, kemudian mengubah semua teks menjadi huruf besar.
b. *Tokenizing*, pemecahan teks berdasarkan karakter spasi.
c. Eliminasi *stopwords*, penghapusan preposisi dan konjungsi.
d. *Stemming*, mengembalikan semua kata menjadi kata dasar.
e. Pembobotan kata menggunakan *tf*idf*, penghapusan kata yang memiliki bobot kecil.

4.2.2 Ekstraksi Topik Dokumen

Ekstraksi Topik Dokumen menggunakan LDA (*Latent Dirichlet Allocation*) membutuhkan beberapa data, yaitu Dokumen, Kamus Kata(V), banyak topik(K), parameter multinomial (β), parameter dirichlet (α), dan banyak iterasi (I). Kamus Kata dibentuk dari ekstraksi semua *term* pada semua Dokumen yang akan diolah.

Pada proses LDA setiap kata pada tiap Dokumen diberi topik secara random, kemudian pada setiap iterasi dilakukan pemutakhiran topik dengan teknik Gibbs Sampling dengan memperhatikan nilai parameter α, β, V, dan K, sehingga diperoleh matriks Phi dan Theta. Dari matriks Phi dapat diketahui *term* dominan atau term yang mempunyai probabilitas tertinggi pada suatu topik seperti pada Tabel 4.3, dan matriks Theta dapat diketahui topik yang dominan dari suatu Dokumen.

4.2.3 Pembentukan *Co-Authorship Graph*

Pembentukan *Co-Authorship Graph* membutuhkan informasi Penulis sebagai node, dan bobot kolaborasi dan bobot kesamaan topik sebagai bobot *edge*. Bobot *edge* diperoleh dengan menggunakan persamaan (7). Bobot kolaborasi didapat dari nilai Jaccard antar penulis atau rasio banyaknya irisan Dokumen yang ditulis oleh

Penulis pertama dan Penulis kedua, dengan banyaknya gabungan Dokumen yang ditulis oleh Penulis pertama dan Penulis kedua.

Bobot kesamaan topik didapat dari representasi matriks Theta dan Penulis Dokumen, yang didapatkan dengan cara normalisasi dari jumlah setiap probabilitas topik terhadap dokumen untuk setiap Dokumen yang ditulis oleh Penulis, sehingga didapatkan probabilitas topik terhadap Penulis (Tabel 4.5).

Probabilitas topik terhadap Penulis ini merepresentasikan komposisi topik dari masing-masing penulis. Bu Diana Purwitasari, 3 topik tertingginya adalah Topik-5 (25,86%), Topik-4 (12,26%) dan Topik-2 (10,14%). Topik-5 merepresentasikan topik Pengolahan Dokumen, Topik-4 merepresentasikan Data Mining, dan Topik-2 merepresentasikan topik Semantik, topik-topik tersebut merepresentasikan topik utama dari Dokumen-Dokumen yang ditulis oleh bu Diana Purwitasari. Seperti pada Tabel A.1, dari 42 Dokumen yang ditulis oleh bu Diana Purwitasari, terdiri dari 14 Dokumen dengan Topik-5 sebagai topik dominan, 10 Dokumen dengan Topik-0 sebagai topik dominan, dan 4 Dokumen dengan Topik-8 sebagai topik dominan.

Dalam penentuan topik penulis ini juga dipengaruhi oleh banyaknya Dokumen (Tabel 4,1) dan variasi topik dominan dari masing-masing Penulis (Tabel 4.4). Bu Chastine Fatichah hanya memiliki 2 Dokumen yang memiliki topik dominan berbeda, sehingga topik Penulis dari bu Chastine Fatichah probabilitasnya hampir merata (Tabel 4.6), dimana besar probabilitas yang tertinggi hanya 16,28% pada Topik-9, tertinggi ke-2 sebesar 16,18% pada Topik-6. Akan tetapi apabila Dokumen yang ditulis memiliki kesamaan topik dominan maka probabilitas penulisnya akan memiliki komposisi yang baik, dimana terdapat topik yang memiliki perbedaan yang signifikan, seperti pada Dokumen yang ditulis oleh pak Radityo Anggoro dimana 2 dari 3 Dokumen yang ditulis oleh beliau memiliki topik dominan Topik-10 (Tabel 4.7) dengan probabilitas sebesar 29,55% dan 28,83%. Berdasarkan kejadian tersebut, maka untuk mendapatkan probabilitas yang dapat mewakili keadaan yang sebenarnya, maka banyak Dokumen yang ditulis oleh Penulis setidaknya sebanyak 5% dari keseluruhan Dokumen yang diolah, dan

Dokumen yang ditulis memiliki topik dominan yang sama setidaknya 70% dari total Dokumen yang ditulis.

Tabel 4.1 Penulis Dokumen pada Jurusan Teknik Informatika

Indeks	Nama Penulis	Banyak Dokumen
0	F.X. Arunanto	4
1	Siti Rochimah	12
2	Fajar Baskoro	2
3	Riyanarto Sarno	33
4	Arya Yudhi Wijaya	20
5	Henning Titi Ciptaningtyas	15
6	Misbakhul Munir Irfan S	1
7	Agus Zainal Arifin	44
8	Sarwosri	29
9	Diana Purwitasari	42
10	Isye Arieshanti	25
11	Rully Sulaiman	65
12	Yudhi Purwananto	51
13	Bilqis Amaliah	26
14	Imam Kuswardayan	33
15	Radityo Anggoro	3
16	Darlis Herumurti	4
17	R. V. Hari Ginardi	4
18	Ahmad Saikhu	21
19	Supeno Djanali	21
20	Ary Mazharuddin Shiddiqi	74
21	Chastine Fatichah	2
22	Muchammad Husni	37
23	Anny Yuniarti	47
24	Waskitho Wibisono	24
25	Handayani Tjandrasa	44
26	Daniel Oranova Siahaan	39
27	Royyana Muslim Ijtihadie	4
28	Umi Laili Yuhana	52
29	Joko Lianto Buliali	14
30	Nanik Suciati	27
31	Dwi Sunaryono	44
32	Suhadi Lili	12
33	Victor Hariadi	7
34	Wahyu Suadi	64

Tabel 4.2 Representasi Data Dokumen

ID_Dokumen	Dokumen	Penulis
5101109048	penentuan status bantu anak asuh pena bangsa menggunakan metode neuro fuzzy sebagai program unggulan program beasiswa anak asuh peduli anak bangsa yayasan dana sosial al falah pena bangsa ydsf memiliki kendala ydsf masih menggunakan cara manual dalam menentukan status bantu anak subyektifitas surveyor di lapangan masih mempengaruhi pemberian status bantu pada anak oleh karena itu diperlukan suatu sistem yang dapat mengklasifikasikan status bantu anak asuh dalam tugas akhir ini akan dibahas pengklasifikasian status bantu dengan benar pengenalan status bantu dilakukan dalam dua tahap tahap pertama adalah melaukukan fuzzifikasi input data sehingga siap digunakan untuk proses pelatihan maupun proses pengenalan pada neural network pada tahap kedua input yang sudah difuzzifikasi diproses dengan neural network arsitektur jaringan yang digunakan adalah backpropagation dengan sebuah input layer 25 input unit sebuah hidden layer hidden unit dapat disesuaikan dan sebuah output layer 3 output unit dari hasil uji coba terhadap jaringan syaraf yang telah terbentuk untuk mengklasifikasikan status bantu maka didapatkan suatu kesimpulan jumlah hidden unit learning rate nilai toleransi error dan fungsi aktivasi mana yang memiliki nilai akurasi paling baik untuk mengklasifikasi status bantu sebagai hasil dari uji coba didapatkan suatu hasil analisis bahwa dari jumlah hidden unit nilai learning rate toleransi error dan fungsi aktivasi maka yang paling baik digunakan untuk mengklasifikasikan status bantu adalah fungsi aktivasi sigmoid biner dengan menggunakan hidden unit 100 momentum 0 2 learning rate 0 01 nilai toleransi error 1e 13 didapatkan tingkat akurasi sebesar 98 70 pada tahap pelatihan dan tingkat akurasi sebesar 93 63 pada tahap pengenalan	Yudhi Purwananto Ahmad Saikhu

Tabel 4.3 Sepuluh kata dominan pada setiap topik pada Jurusan Teknik Informatika, FTIf, ITS (K=12, α = 2 dan β = 0,05)

Topik ke-0	(%)	Topik ke-1	(%)	Topik ke-2	(%)	Topik ke-3	(%)
bisnis	4,62	segmen	7,56	web	10,61	lokasi	4,10
manajemen	3,17	detik	4,00	beritahu	4,43	posisi	3,87
lunak	2,19	nokia	2,70	interop	3,76	induk	3,13
orientasi	2,04	warna	2,40	aktor	3,57	petabyte	2,89
architect	1,98	fine	2,29	rekonstruksi	2,87	bayar	2,28
domain	1,88	return	2,21	size	2,47	social	1,79
desain	1,83	transisi	1,97	kontrol	2,14	wisata	1,73
lapor	1,72	tiada	1,87	gpgpu	1,63	mapserver	1,68
produk	1,69	buta	1,83	mail	1,60	tea	1,49
integrasi	1,56	estate	1,54	jam	1,55	plan	1,48

Topik ke-4	(%)	Topik ke-5	(%)	Topik ke-6	(%)	Topik ke-7	(%)
klasifikasi	4,80	dokumen	7,94	*cluster*	6,17	main	7,99
gabor	4,52	bahaya	2,80	optimal	4,99	aneka	4,53
ekualisasi	2,04	teks	2,79	klasik	3,78	mobile	4,50
kansei	2,02	luncur	2,52	solusi	3,25	gamma	4,06
vector	1,77	topik	2,23	preetham	2,59	jam	3,64
device	1,70	keluh	2,20	tree	2,27	suara	2,27
neural	1,68	spline	1,84	australia	1,59	telepon	2,18
terapi	1,47	bom	1,56	kemoterapi	1,51	opnet	1,78
kerap	1,44	semen	1,55	deblur	1,51	user	1,70
knowledge	1,44	kampus	1,42	keluh	1,49	smu	1,59

Topik ke-8	(%)	Topik ke-9	(%)	Topik ke-10	(%)	Topik ke-11	(%)
file	5,61	gerak	5,31	jaring	10,82	ruang	2,93
pesan	4,07	obyek	4,41	protokol	3,95	ditektor	2,67
komputasi	4,00	baterai	3,54	no*graph*is	3,03	gilir	2,02
median	3,69	robot	3,01	kondisi	2,82	agility	1,84
video	2,70	realtime	2,49	route	2,05	falah	1,81
entropi	2,32	simulasi	2,21	palsu	1,97	diimput	1,59
streambase	1,86	keping	1,77	modelingpartial	1,91	color	1,57
kolmogrov	1,67	koneksi	1,69	rute	1,85	rumah	1,54
step	1,66	sentence	1,68	voip	1,64	profesi	1,53
koneksi	1,54	momen	1,63	gameplay	1,60	inferensi	1,49

Tabel 4.4 Matriks probabilitas topik terhadap Penulis Dokumen di Jurusan Teknik Infromatika, FTIf, ITS

Indeks	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
0	4,76	2,92	19,30	6,16	5,32	8,83	8,83	3,15	21,36	3,70	7,42	8,27
1	10,69	3,14	10,80	9,86	3,57	16,51	4,29	18,73	3,67	10,07	4,04	4,63
2	14,47	4,62	24,98	5,25	4,47	7,19	3,34	7,38	13,19	6,41	4,06	4,63
3	49,88	2,89	5,67	3,55	3,21	10,32	4,07	3,40	3,26	5,06	3,79	4,90
4	3,65	34,32	6,13	7,24	10,49	4,80	5,27	3,68	7,59	6,92	5,39	4,54
5	4,10	4,60	10,56	5,66	7,55	5,15	4,20	17,10	7,29	7,42	18,62	7,74
6	5,07	31,53	4,68	4,15	5,91	6,13	5,48	3,46	6,34	3,62	4,29	19,36
7	3,76	22,73	3,66	7,73	12,00	12,38	12,50	3,66	3,56	6,05	3,94	8,03
8	9,71	3,14	14,20	9,28	4,06	18,34	6,91	8,94	5,56	6,25	3,85	9,76
9	4,79	8,26	10,14	8,53	12,26	25,83	7,86	3,18	3,59	5,27	3,32	6,97
10	4,08	11,32	5,61	5,02	22,44	5,64	20,45	3,79	4,00	5,21	4,67	7,77
11	4,85	16,79	3,75	4,76	14,77	4,29	15,48	5,94	6,35	9,02	4,72	9,28
12	4,77	15,27	4,21	4,96	13,10	4,42	18,22	6,15	7,01	8,54	5,53	7,82
13	4,68	22,01	3,10	5,48	18,25	3,48	9,87	3,46	3,93	5,28	5,01	15,45
14	5,46	3,64	6,85	8,59	3,78	5,07	3,71	32,00	4,06	18,82	3,77	4,26
15	3,72	3,86	19,31	4,37	3,38	3,51	4,03	9,20	8,90	9,86	24,91	4,95
16	5,05	16,53	4,29	4,15	16,08	6,61	6,66	3,37	4,60	7,96	6,30	18,39
17	3,55	15,34	8,65	2,90	28,98	4,41	19,38	2,57	2,75	3,97	3,61	3,89
18	8,74	4,78	4,11	7,78	18,62	7,66	27,40	2,97	3,48	4,81	3,74	5,91
19	4,42	2,87	3,74	3,48	4,24	3,28	7,47	3,70	9,80	4,48	47,48	5,04
20	5,39	3,54	12,43	11,66	3,64	4,60	4,59	13,05	14,34	6,77	12,10	7,89
21	6,24	8,51	5,85	5,13	13,24	8,86	16,18	4,87	5,27	16,28	5,03	4,54
22	4,54	3,74	9,27	6,75	3,49	4,13	4,11	8,25	15,75	6,80	25,61	7,56
23	4,04	26,23	5,12	5,92	13,12	6,36	10,45	4,04	4,55	7,56	3,27	9,34
24	3,37	4,97	6,48	10,25	7,18	4,34	6,73	14,45	11,37	6,34	12,07	12,44
25	4,21	37,95	2,94	3,54	14,50	4,12	9,29	3,50	4,78	5,19	4,57	5,41
26	8,37	3,90	5,45	6,21	5,52	28,64	14,28	3,38	3,76	10,43	3,03	7,03
27	7,02	3,98	24,42	4,15	3,07	4,11	3,34	7,26	16,23	3,83	16,11	6,48
28	9,24	3,22	12,16	14,48	4,11	17,36	5,54	7,59	4,67	9,94	4,25	7,44
29	6,26	3,59	13,23	8,82	9,77	5,92	15,28	4,27	6,55	13,62	4,32	8,37
30	4,26	23,26	3,74	8,57	15,33	5,43	7,94	4,90	8,33	8,02	4,27	5,95
31	22,55	3,11	15,16	7,10	3,51	8,47	4,54	10,65	4,04	11,60	3,95	5,32
32	12,42	3,28	3,96	4,56	4,77	8,34	3,80	34,75	4,30	10,40	4,65	4,77
33	12,25	8,78	6,74	7,22	9,19	3,34	16,80	6,37	5,59	5,10	6,92	11,69
34	5,51	4,34	13,11	6,08	4,34	4,91	4,55	11,14	17,51	9,58	13,48	5,44

Tabel 4.5 Representasi Dokumen setelah praproses

status anak asuh pena bangsa neuro unggul beasiswa anak asuh peduli anak bangsa yayasan dana social falah pena bangsa manual status anak subjek survey lapang status anak status anak asuh bahas status status neural neural architect backpropagasi layer hide layer hide output layer output syaraf status hide learn rate toleransi status hide learn rate toleransi status sigmoid biner hide momen learn rate toleransi

Tabel 4.6 Probabilitas topik Dokumen yang ditulis oleh bu Chastine.

ID_Dokumen	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
5104100039	9,06	12,64	4,83	4,90	6,46	5,33	5,30	6,12	7,37	27,46	4,68	5,86
5105100153	3,42	4,38	6,87	5,36	20,02	12,40	27,05	3,62	3,17	5,10	5,39	3,22

Tabel 4.7 Probabilitas topik Dokumen yang ditulis oleh pak Radityo.

ID_Dokumen	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
5103109040	4,87	3,42	7,39	4,48	3,81	4,33	3,94	7,06	5,49	21,85	29,55	3,81
5104100126	2,43	2,49	27,23	5,32	2,18	2,85	4,01	16,47	12,74	3,64	16,36	4,27
5105100164	3,86	5,66	23,3	3,31	4,14	3,35	4,14	4,08	8,48	4,09	28,83	6,76

Setelah bobot kolaborasi dan bobot kesamaan topik tersedia, berikutnya dihitung bobot edge menggunakan persamaan (7). Pada persamaan (7) terdapat konstanta *c*, yang digunakan untuk rasio antara bobot kolaborasi dan bobot kesamaan topik. Nilai *c* minimum atau *c* = 0,1, merepresentasikan bahwa bobot edge lebih memprioritaskan kolaborasi daripada kesamaan topik, dan sebaliknya jika nilai *c* maksimal atau *c* = 0,9, maka bobot edge lebih memprioritaskan kesamaan topik daripada kolaborasi.

4.2.4 Pengelompokan dengan *Hypergraph-Partitioning*

Node-node pada *graph* yang sudah terbentuk dikelompokan menggunakan *Hypergraph-Partitioning*. Pada tahap coarsening, semua node dibagi menjadi 2 kelompok atau lebih dari 2 jika antar kelompok yang terbentuk tidak dapat digabungkan, dimana tiap tahap penggabungan pada tiap level disimpan. Tahap

Balancing, meminimalisasi banyaknya *edge* yang terpotong dari 2 kelompok hasil coarsening. *Uncoarsening* , dengan menggunakan informasi iterasi pada proses coarsening, dilakukan proses *Balancing* untuk setiap pasang kelompok yang pada iterasi selanjutnya bergabung.

Hasil pengelompokan adalah hasil penelusuran iterasi *uncoarsening* dimulai dari kelompok induk ke cabang-cabangnya, jika banyak anggota kelompok lebih kecil atau sama dengan banyak anggota yang diinginkan, maka kelompok tersebut disimpan sebagai hasil pengelompokan.

4.3 Uji coba 1 : Pengaruh banyak topik dan nilai *c* terhadap kualitas pengelompokan satu Jurusan.

Penentuan banyak topik pada proses ekstraksi topik menentukan dimensi dari data Penulis, dimana probabilitas topik penulis adalah representasi dari data Penulis dan matriks Theta atau probabilitas topik terhadap Dokumen, sehingga semakin besar nilai *K* (banyak topik) maka semakin besar dimensinya. Nilai *c* pada perhitungan bobot edge mempengaruhi graph yang terbentuk, apakah lebih memprioritaskan kolaborasi atau kesamaan topik.

Uji coba ini bertujuan untuk mengetahui nilai *K* dan *c* yang memberikan hasil pengelompokan Penulis paling baik, hasil uji coba dilihat dari nilai *Average Silhouette Width* (ASW) masing-masing pengelompokan. Nilai *K* untuk pengujian ini adalah 12, 15, 18, 21 dan 24, sementara *c* adalah 0,1 sampai 0,9.

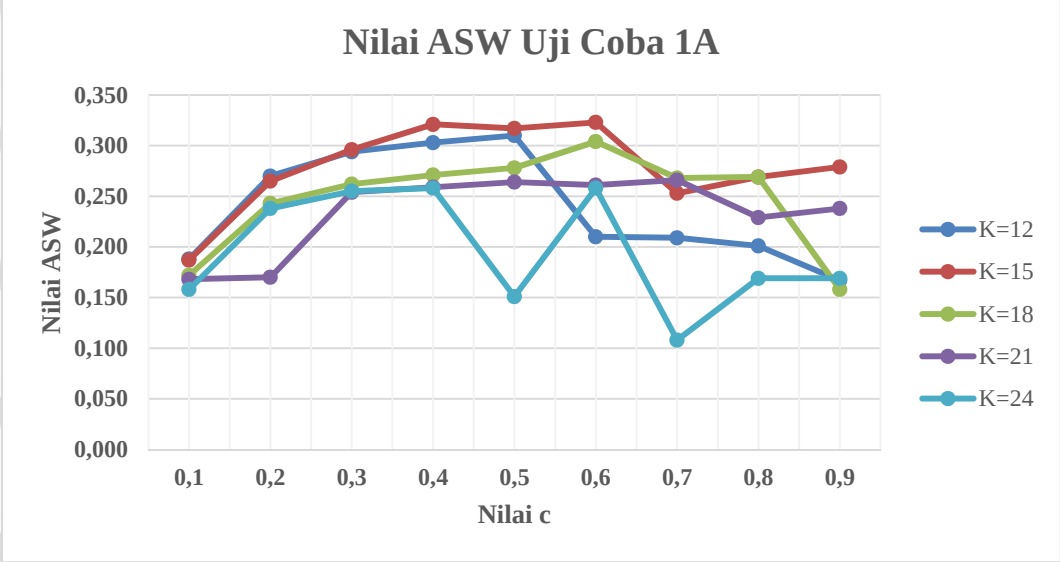
Parameter Latent Dirichlet Allocation (LDA) yang digunakan pada pengujian ini adalah $\alpha = 2$, $\beta = 0{,}05$ dan $I = 2000$. Parameter untuk *Hypergraph-Partitioning*, banyak maksimal anggota kelompok adalah 9 Penulis.

4.3.1. Uji Coba 1A

Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Teknik Informatika, FTIf, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 639 Dokumen dan 35 Penulis. Pada pengujian ini Penulis pada Jurusan Teknik Informatika terkelompok menjadi 5 kelompok.

Tabel 4.8 Nilai *Average Silhouette Width*, pada uji coba 1A.

Nilai *c*	*Average Silhouette Width*				
	***K*=12**	***K*=15**	***K*=18**	***K*=21**	***K*=24**
0,1	0,188	0,187	0,172	0,168	0,158
0,2	0,270	0,265	0,243	0,170	0,238
0,3	0,294	0,296	0,262	0,254	0,255
0,4	0,303	0,321	0,271	0,259	**0,258**
0,5	**0,310**	0,317	0,278	0,264	0,151
0,6	0,210	**0,323**	**0,304**	0,261	**0,258**
0,7	0,209	0,253	0,268	**0,266**	0,108
0,8	0,201	0,269	0,269	0,229	0,169
0,9	0,167	0,279	0,158	0,238	0,169



Gambar 4.1 Grafik nilai ASW uji coba 1A

Tabel 4.8 menunjukkan nilai ASW dari masing-masing pengelompokan, dimana nilai ASW tertinggi diperoleh saat c = 0,6 dan K = 15, hal ini menunjukkan bahwa struktur dari pengelompokan tersebut lebih baik daripada hasil pengelompokkan lainnya. Secara keseluruhan berdasarkan nilai ASW pengelompokkan dengan nilai K = 12, 15 dan 18 memiliki struktur yang lebih baik daripada K = 21 dan 24, selain itu tidak didapatkan kepastian tentang nilai c yang tepat untuk mendapatkan hasil yang terbaik karena pada K = 12 ASW terbaik didapat dari c = 0,5, pada K = 15 dan 18 ASW terbaik didapat dari c = 0,6, dan pada

K = 21 ASW terbaik didapat dari *c* = 0,7 dari *K* = 12 sampai 21 didapat kecenderungan nilai ASW terbaik didapat seiring dengan kenaikan nilai c, tetapi pada *K*=24 ASW terbaik didapat dari *c* = 0,4 dan *c* = 0,6.

Berdasarkan grafik nilai ASW pada Gambar 4.1, *K* = 12 memiliki keteraturan nilai ASW yang baik, dimana pada *c* = 0,1 sampai 0,5 naik dan *c* = 0,5 sampai 0,9 menurun. Sementara grafik yang tidak beraturan pada *K* = 24, dimana grafik nilai ASW-nya naik turun.

Pada pengujian ini nilai *c* terbaik adalah pada rentang 0,4 sampai 0,7 dan nilai *K* yang terbaik adalah antara 12 sampai 18.

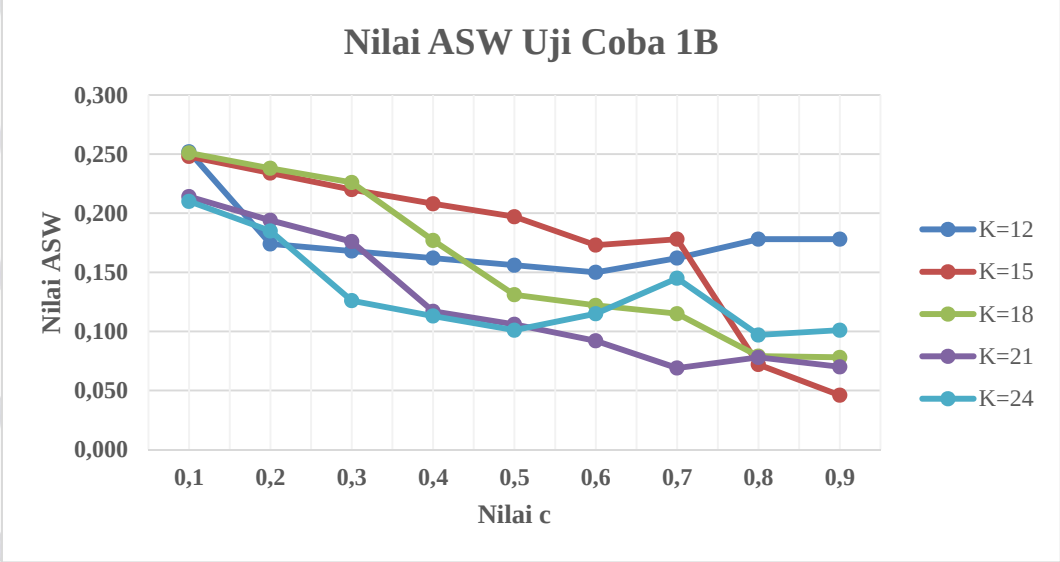
4.3.2. Uji Coba 1B

Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Sistem Informasi, FTIf, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 329 Dokumen dan 22 Penulis. Pada pengujian ini Penulis pada Jurusan Sistem Informasi terkelompok menjadi 3 kelompok.

Tabel 4.9 menunjukkan nilai ASW dari masing-masing pengelompokkan, Kelompok yang menggunakan nilai *K* = 12 dan *c* = 0,1 menjadi kelompok yang paling baik dengan nilai ASW tertinggi sebesar 0,252. Secara keseluruhan kelompok yang terbaik selalu didapatkan dari nilai *c* = 0,1, berbeda dengan uji coba 1A pada Jurusan Teknik Informatika. Berdasarkan grafik nilai ASW uji coba 1B (Gambar 4.2) menunjukkan sebuah keteraturan dimana semakin besar nilai c, maka nilai ASW semakin menurun. Pada perhitungan bobot edge nilai bobot kolaborasi adalah sama untuk semua nilai K, tetapi nilai kesamaan topik dipengaruhi oleh banyaknya topik, sehingga perlu diketahui perubahan nilai kesamaan topik dari *K* = 12 ke *K* = 15, *K* = 15 ke *K* = 18, dan seterusnya apakah naik atau turun. Tabel 4.10 menunjukkan banyaknya perubahan nilai kesamaan topik, nampak bahwa pola perubahan yang terjadi lebih cenderung naik-turun-naik-turun dan distribusi polanya merata, sehingga bobot kesamaan topik tiap nilai *K* tidak terlalu berbeda. Tidak seperti pada Jurusan Teknik Informatika dimana pola naik-turun-naik-naik sangat mendominasi hingga 20,30% dari pola yang ada.

Tabel 4.9 Nilai *Average Silhouette Width*, pada uji coba 1B.

Nilai *c*	Average Silhouette Width				
	***K*=12**	***K*=15**	***K*=18**	***K*=21**	***K*=24**
0,1	**0,252**	**0,248**	**0,251**	**0,214**	**0,210**
0,2	0,174	0,234	0,238	0,194	0,185
0,3	0,168	0,220	0,226	0,176	0,126
0,4	0,162	0,208	0,177	0,117	0,113
0,5	0,156	0,197	0,131	0,106	0,101
0,6	0,150	0,173	0,122	0,092	0,115
0,7	0,162	0,178	0,115	0,069	0,145
0,8	0,178	0,072	0,079	0,078	0,097
0,9	0,178	0,046	0,078	0,070	0,101



Gambar 4.2 Grafik nilai ASW uji coba 1B

Tabel 4.10 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 1A dan 1B.

Perubahan nilai dari *K* = a ke *K* = B				Prosentase	
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	**T.Informatika**	**S.Informasi**
turun	turun	turun	naik	2,88	1,73
turun	turun	naik	turun	5,58	12,12
turun	turun	naik	naik	9,14	10,39
turun	naik	turun	turun	1,35	5,63
turun	naik	turun	naik	6,26	1,73
turun	naik	naik	turun	10,83	12,99

Perubahan nilai dari *K* = a ke *K* = b				Prosentase	
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	**T.Informatika**	**S.Informasi**
turun	naik	naik	naik	12,52	3,03
naik	turun	turun	turun	0,68	0,87
naik	turun	turun	naik	3,05	3,90
naik	turun	naik	turun	7,95	**17,75**
naik	turun	naik	naik	**20,30**	16,45
naik	naik	turun	turun	0,17	0,87
naik	naik	turun	naik	4,91	1,73
naik	naik	naik	turun	4,91	6,93
naik	naik	naik	naik	9,48	3,90

4.3.3. Uji Coba 1C

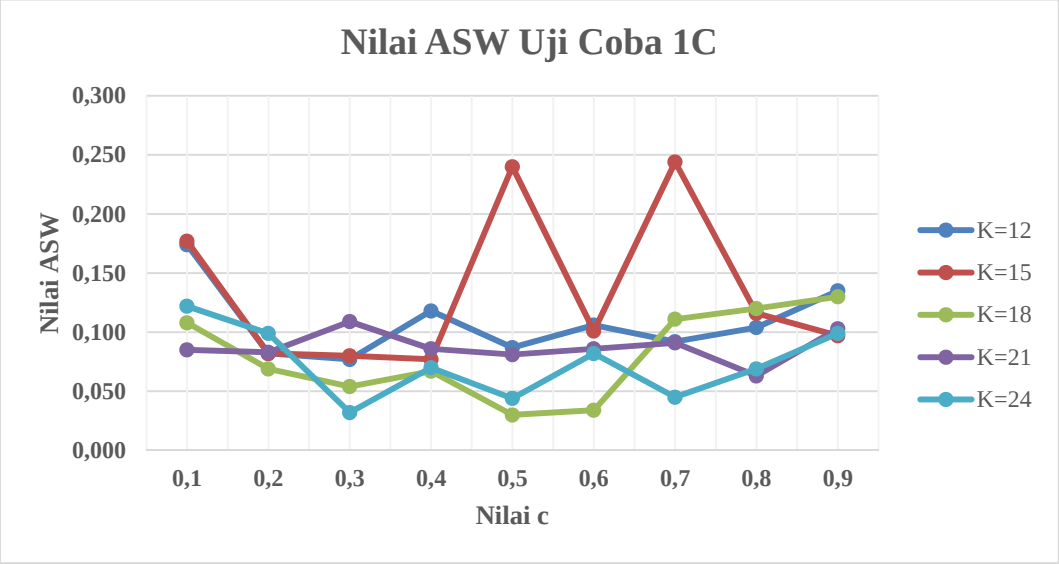
Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Teknik Industri, FTI, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 581 Dokumen dan 34 Penulis. Pada pengujian ini Penulis pada Jurusan Teknik Industri terkelompok menjadi 5 kelompok.

Pada pengujian ini nilai ASW terbaik sebesar 0,244 didapat dari nilai *K* = 15 dan *c* = 0,7 seperti pada Tabel 4.11. Secara keseluruhan didapatkan pola dimana saat nilai *K* antara 12 dan 18 nilai ASW terbaik didapat pada nilai *c* yang semakin besar dan saat *K* antara 18 dan 24 nilai ASW terbaik didapat pada nilai *c* yang semakin kecil.

Tabel 4.11 Nilai *Average Silhouette Width*, pada uji coba 1C.

Nilai *c*	Average Silhouette Width				
	***K*=12**	***K*=15**	***K*=18**	***K*=21**	***K*=24**
0,1	**0,174**	0,177	0,108	0,085	**0,122**
0,2	0,082	0,082	0,069	0,083	0,099
0,3	0,077	0,080	0,054	**0,109**	0,032
0,4	0,118	0,077	0,067	0,086	0,070
0,5	0,087	0,240	0,030	0,081	0,044
0,6	0,106	0,101	0,034	0,086	0,082
0,7	0,092	**0,244**	0,111	0,091	0,045
0,8	0,104	0,116	0,120	0,063	0,069
0,9	0,135	0,097	**0,130**	0,103	0,099

Sedangkan berdasarkan grafil nilai ASW pada pengujian ini (Gambar 4.3) nampak ketidakteraturan nilai ASW baik berdasarkan nilai *K* ataupun nilai *c*.



Gambar 4.3 Grafik nilai ASW uji coba 1C

Tabel 4.12 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 1C

Perubahan nilai dari *K* = a ke *K* = b				**Prosentase**
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	
turun	turun	turun	naik	2,33
turun	turun	naik	turun	6,44
turun	turun	naik	naik	9,30
turun	naik	turun	turun	4,11
turun	naik	turun	naik	12,88
turun	naik	naik	turun	9,30
turun	naik	naik	naik	**17,17**
naik	turun	turun	turun	0,72
naik	turun	turun	naik	2,33
naik	turun	naik	turun	6,62
naik	turun	naik	naik	13,24
naik	naik	turun	turun	1,25
naik	naik	turun	naik	1,97
naik	naik	naik	turun	3,58
naik	naik	naik	naik	8,77

Berdasarkan pola prosentase perubahan kesamaan topik Penulis antar nilai K, Tabel 4.12 didapati pola turun-naik-naik-naik memiliki prosentase terbesar, tetapi distribusi pola perubahannya merata, dimana selisih prosentasenya tidak terlalu tinggi.

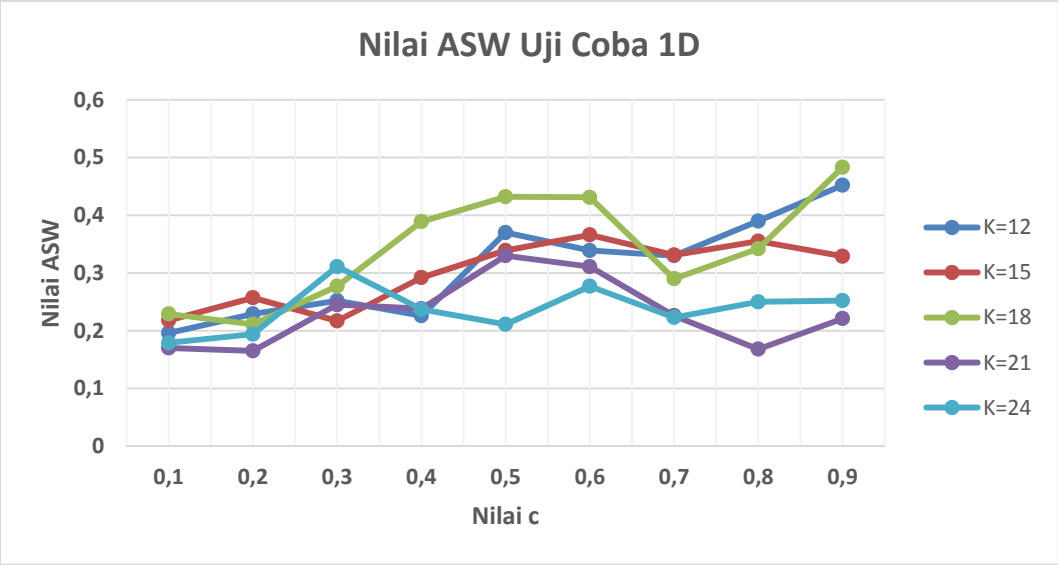
4.3.4. Uji Coba 1D

Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Teknik Elektro, FTI, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 1.544 Dokumen dan 75 Penulis. Pada pengujian ini Penulis pada Jurusan Teknik Elektro terkelompok menjadi 10 kelompok.

Tabel 4.13 Nilai *Average Silhouette Width*, pada uji coba 1D.

Nilai *c*	**Average Silhouette Width**				
	***K*=12**	***K*=15**	***K*=18**	***K*=21**	***K*=24**
0,1	0,196	0,218	0,229	0,170	0,179
0,2	0,229	0,257	0,211	0,165	0,194
0,3	0,252	0,217	0,277	0,245	**0,311**
0,4	0,226	0,292	0,389	0,238	0,237
0,5	0,370	0,339	0,432	**0,330**	0,211
0,6	0,339	**0,366**	0,431	0,311	0,277
0,7	0,330	0,331	0,290	0,226	0,223
0,8	0,390	0,355	0,342	0,168	0,250
0,9	**0,452**	0,329	**0,483**	0,221	0,252

Dari Tabel 4.13, diketahui bahwa nilai ASW terbaik sebesar 0,483 diperoleh dari nilai K = 18 dan c = 0,9. Secara keseluruhan terdapat 2 nilai K yang memberikan nilai ASW tertinggi yaitu 12 dan 18 dimana nilai ASW terbaiknya memiliki perbedaan yang signifikan dengan 3 nilai K lainnya. Selain itu tidak terdapat pola khusus terkait nilai K dan c yang memberikan nilai ASW terbaik. Pada grafik nilai ASW, Gambar 3.8, K = 12, 15 dan 18 memberikan grafik naik seiring dengan kenaikan nilai c, tetapi K = 21 dan 24 memberikan grafik naik-turun seiring kenaikan nilai c, hal ini memberikan gambaran bahwa pada pengujian ini nilai K = 21 dan 24 tidak disarankan karena memberikan nilai ASW yang kurang baik.



Gambar 4.4 Grafik nilai ASW uji coba 1D

Tabel 4.14 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 1D

Perubahan nilai dari *K* = a ke *K* = b				**Prosentase**
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	
turun	turun	turun	naik	2,01
turun	turun	naik	turun	2,48
turun	turun	naik	naik	3,14
turun	naik	turun	turun	6,09
turun	naik	turun	naik	11,28
turun	naik	naik	turun	6,35
turun	naik	naik	naik	6,42
naik	turun	turun	turun	1,17
naik	turun	turun	naik	5,11
naik	turun	naik	turun	6,17
naik	turun	naik	naik	9,78
naik	naik	turun	turun	5,29
naik	naik	turun	naik	**13,18**
naik	naik	naik	turun	9,78
naik	naik	naik	naik	11,75

Perubahan nilai *K* juga menyebabkan perubahan nilai bobot kesamaan topik Penulis, Tabel 4.14, pola naik-naik-turun-naik dan naik-naik-naik-naik memiliki prosentase tertinggi, akan tetapi berdasarkan nilai ASW, kenaikan nilai bobot

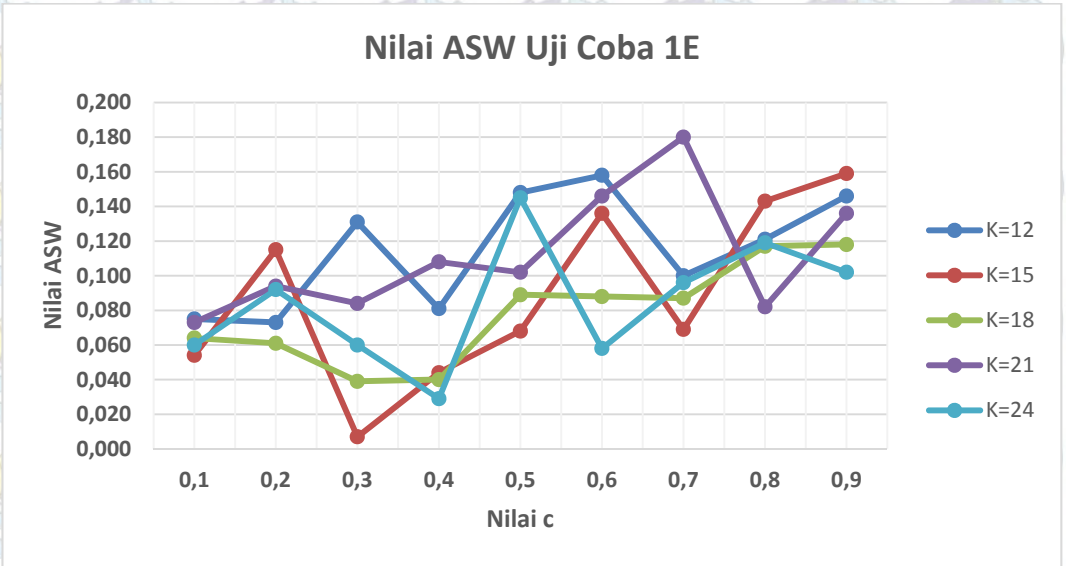
kesamaan topik Penulis ini tidak menjamin dapat memberikan nilai ASW yang baik.

4.3.5. Uji Coba 1E

Data yang digunakan untuk pengujian ini adalah Dokumen dari Jurusan Statistika, FMIPA, ITS Surabaya dari tahun 2006 sampai 2013, sebanyak 615 Dokumen dan 35 Penulis. Pada pengujian ini Penulis pada Jurusan Statistika terkelompok menjadi 5 kelompok.

Tabel 4.15 Nilai *Average Silhouette Width*, pada uji coba 1E.

Nilai *c*	Average Silhouette Width				
	***K*=12**	***K*=15**	***K*=18**	***K*=21**	***K*=24**
0,1	0,075	0,054	0,064	0,073	0,060
0,2	0,073	0,115	0,061	0,094	0,092
0,3	0,131	0,007	0,039	0,084	0,060
0,4	0,081	0,044	0,040	0,108	0,029
0,5	0,148	0,068	0,089	0,102	**0,145**
0,6	**0,158**	0,136	0,088	0,146	0,058
0,7	0,100	0,069	0,087	**0,180**	0,096
0,8	0,121	0,143	0,117	0,082	0,119
0,9	0,146	**0,159**	**0,118**	0,136	0,102



Gambar 4.5 Grafik nilai ASW uji coba 1E

Pada Tabel 4.15 diketahui bahwa nilai ASW tertinggi sebesar 0,180 didapatkan dari nilai *K* = 21 dan *c* = 0,7. Secara keseluruhan nilai ASW terbaik didapat dari nilai *c* lebih dari sama dengan 0,5. Pada pengujian ini nilai ASW masing-masing nilai K, tidak mempunyai keteraturan seperti pada grafik pada Gambar 4.5, setiap kenaikan nilai c, selisih nilai ASWnya sangat besar.

Tabel 4.16 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 1E

Perubahan nilai dari *K* = a ke *K* = b				**Prosentase**
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	
turun	turun	turun	naik	2,71
turun	turun	naik	turun	3,21
turun	turun	naik	naik	2,88
turun	naik	turun	turun	2,03
turun	naik	turun	naik	12,35
turun	naik	naik	turun	5,25
turun	naik	naik	naik	5,92
naik	turun	turun	turun	1,35
naik	turun	turun	naik	15,06
naik	turun	naik	turun	7,61
naik	turun	naik	naik	12,35
naik	naik	turun	turun	4,06
naik	naik	turun	naik	**16,24**
naik	naik	naik	turun	4,91
naik	naik	naik	naik	4,06

Pola naik-naik-turun-naik memiliki probabilitas yang tinggi, Tabel 4.16, akan tetapi sekalipun nilai bobot persamaan topik antar Penulisnya naik, tidak memberikan jaminan bahwa nilai ASWnya juga semakin baik.

4.3.6. Analisa Uji Coba 1

Dari skenario 1A-1E ujicoba pengaruh banyaknya topik (*K*) dan nilai *c* terhadap kualitas pengelompokan satu Jurusan, diperoleh :

a. Tidak terdapat kepastian hubungan antara nilai *c* dan *K* dengan nilai ASW, dimana pada setiap pengujian nilai ASW didapatkan dari nilai *c*

dan *K* yang berbeda-beda. Tetapi secara keseluruhan nilai ASW terbaik sebesar 0,483 didapatkan saat menggunakan data Jurusan Elektro (Tabel 4.17), hal ini menunjukkan bahwa pada hasil pengelompokan Penulis Jurusan Teknik Elektro, jarak antar kelompok Penulis sangat baik tetapi secara struktur pengelompokannya masih lemah. Sementara pengelompokan pada Jurusan Statistika mendapatkan nilai ASW terendah sebesar 0,180, hal ini menunjukkan bahwa pengelompokan Penulis pada Jurusan Statistika, jarak antar kelompok Penulis kurang baik dan memiliki struktur terlemah dibandingkan lima Jurusan lainnya.

b. Perubahan nilai *K* menyebabkan perubahan nilai bobot kesamaan topik antar Penulis, sehingga membentuk pola naik dan turun, pada pengujian ini didapatkan pola perubahan seperti pada tabel 4.18. Perubahan yang memiliki prosentase terbesar adalah naik-turun-naik-naik pada data Penulis Jurusan Teknik Informatika, akan tetapi sekalipun pola yang terjadi adalah naik, nilai ASWnya tidak terlalu terpengaruh oleh kenaikan tersebut.

Tabel 4.17 Nilai c, *K* dan ASW dari Uji Coba 1.

Uji Coba	nilai c	nilai K	ASW
1A	0,6	15	0,323
1B	0,1	12	0,252
1C	0,7	15	0,244
1D	0,9	18	0,483
1E	0,7	21	0,180

Tabel 4.18 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 1

Perubahan nilai dari *K* = a ke *K* = b				Prosentase Uji Coba ke-				
12 ke 15	15 ke 18	18 ke 21	21 ke 24	1A	1B	1C	1D	1E
turun	turun	turun	naik	2,88	1,73	2,33	2,01	2,71
turun	turun	naik	turun	5,58	12,12	6,44	2,48	3,21
turun	turun	naik	naik	9,14	10,39	9,30	3,14	2,88
turun	naik	turun	turun	1,35	5,63	4,11	6,09	2,03
turun	naik	turun	naik	6,26	1,73	12,88	11,28	12,35

Perubahan nilai dari *K* = a ke *K* = B				Prosentase Uji Coba ke-				
12 ke 15	**15 ke 18**	**18 ke 21**	**21 ke 24**	**1A**	**1B**	**1C**	**1D**	**1E**
turun	naik	naik	turun	10,83	12,99	9,30	6,35	5,25
turun	naik	naik	naik	12,52	3,03	**17,17**	6,42	5,92
naik	turun	turun	turun	0,68	0,87	0,72	1,17	1,35
naik	turun	turun	naik	3,05	3,90	2,33	5,11	15,06
naik	turun	naik	turun	7,95	**17,75**	6,62	6,17	7,61
naik	turun	naik	naik	**20,30**	16,45	13,24	9,78	12,35
naik	naik	turun	turun	0,17	0,87	1,25	5,29	4,06
naik	naik	turun	naik	4,91	1,73	1,97	**13,18**	**16,24**
naik	naik	naik	turun	4,91	6,93	3,58	9,78	4,91
naik	naik	naik	naik	9,48	3,90	8,77	11,75	4,06

4.4 Uji coba 2 : Pengaruh banyak topik dan nilai *c* terhadap kualitas pengelompokan beberapa Jurusan.

Penentuan banyak topik pada proses ekstraksi topik menentukan dimensi dari data Penulis, dimana probabilitas topik penulis adalah representasi dari data Penulis dan matriks Theta atau probabilitas topik terhadap Dokumen, sehingga semakin besar nilai *K* (banyak topik) maka semakin besar dimensinya. Nilai *c* pada perhitungan bobot edge mempengaruhi graph yang terbentuk, apakah lebih memprioritaskan kolaborasi atau kesamaan topik.

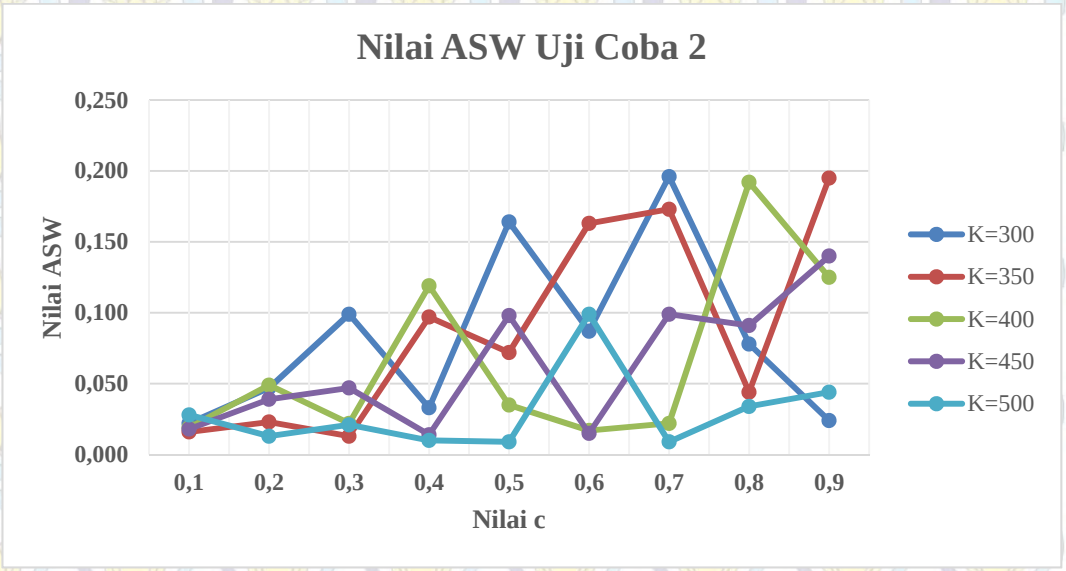
Uji coba ini bertujuan untuk mengetahui nilai *K* dan *c* yang memberikan hasil pengelompokan Penulis paling baik, dimana data yang digunakan berasal dari beberapa Jurusan. Hasil uji coba dilihat dari nilai *Average Silhouette Width* (ASW) masing-masing pengelompokan. Nilai *c* untuk pengujian ini 0,1 sampai 0,9.

Parameter Latent Dirichlet Allocation (LDA) yang digunakan pada pengujian ini adalah $\alpha = 2$, $\beta = 0{,}05$ dan $I = 2000$. Parameter untuk *Hypergraph-Partitioning*, banyak maksimal anggota kelompok adalah 9 Penulis.

Data yang digunakan untuk pengujian ini adalah Dokumen dari 22 Jurusan di ITS Surabaya (Tabel 3.1) dari tahun 2006 sampai 2013, sebanyak 10.722 Dokumen dan 751 Penulis. Pada pengujian ini Penulis dari dari 22 Jurusan tersebut terkelompok menjadi 94 kelompok.

Tabel 4.19 Nilai ASW pada Uji Coba 2

N_Topik	Average Silhouette Width				
	***K*=300**	***K*=350**	***K*=400**	***K*=450**	***K*=500**
0,1	0,022	0,016	0,019	0,018	0,028
0,2	0,047	0,023	0,049	0,039	0,013
0,3	0,099	0,013	0,022	0,047	0,021
0,4	0,033	0,097	0,119	0,014	0,010
0,5	0,164	0,072	0,035	0,098	0,009
0,6	0,087	0,163	0,017	0,015	**0,099**
0,7	**0,196**	0,173	0,022	0,099	0,009
0,8	0,078	0,044	**0,192**	0,091	0,034
0,9	0,024	**0,195**	0,125	**0,140**	0,044



Gambar 4.6 Grafik nilai ASW Uji Coba 2

Pada Tabel 4.19 diketahui bahwa nilai ASW tertinggi sebesar 0,196 diperoleh dari nilaik *K* = 300 dan *c* = 0,7. Secara keseluruhan nilai ASW tertinggi diperoleh dengan nilai *c* lebih dari 0,6, hal ini mengindikasikan bahwa pengelompokan dengan memprioritaskan bobot kesamaan topik antar Penulis menghasilkan pengelompokkan yang baik, hal ini dapat dilihat juga dari grafik nilai ASW (Gambar 4.6) dimana grafik cenderung naik seiring dengan kenaikan nilai *c*.

Pada Tabel 4.20 pola perubahan nilai bobot kesamaan topik yang memiliki prosentase tertinggi adalah naik-naik-turun-naik dan naik- naik- naik- naik, hal ini

mengindikasikan bahwa bobot kesamaan topik antar Penulis akan semakin baik seiring semakin tingginya nilai K, akan tetapi perubahan ini tidak berpengaruh pada nilai ASW dan justru nilai ASW-nya semakin menurun seiring kenaikan nilai *K*.

Tabel 4.20 Prosentase pola perubahan nilai kesamaan topik antar nilai *K* secara beruntun pada uji coba 2

Perubahan nilai dari *K* = a ke *K* = b				**Prosentase**
300 ke 350	**350 ke 400**	**400 ke 450**	**450 ke 500**	
turun	turun	turun	naik	0,50
turun	turun	naik	turun	2,23
turun	turun	naik	naik	0,37
turun	naik	turun	turun	4,27
turun	naik	turun	naik	3,76
turun	naik	naik	turun	4,55
turun	naik	naik	naik	3,25
naik	turun	turun	turun	2,51
naik	turun	turun	naik	5,80
naik	turun	naik	turun	12,01
naik	turun	naik	naik	7,26
naik	naik	turun	turun	5,99
naik	naik	turun	naik	**18,00**
naik	naik	naik	turun	11,92
naik	naik	naik	naik	17,57

4.5 Uji coba 3 : Pengaruh banyak topik dan nilai *c* terhadap validitas pengelompokan.

Penentuan banyak topik pada proses ekstraksi topik menentukan dimensi dari data Penulis, dimana probabilitas topik penulis adalah representasi dari data Penulis dan matriks Theta atau probabilitas topik terhadap Dokumen, sehingga semakin besar nilai *K* (banyak topik) maka semakin besar dimensinya. Nilai *c* pada perhitungan bobot edge mempengaruhi graph yang terbentuk, apakah lebih memprioritaskan kolaborasi atau kesamaan topik.

Uji coba ini bertujuan untuk mengetahui nilai *K* dan *c* yang memberikan hasil pengelompokan Penulis paling baik, hasil uji coba dilihat dari nilai *Entropy* masing-masing pengelompokan dengan data Peneliti dari :

a. Penelitian Dosen tahun 2012
b. Penelitian Dosen tahun 2013
c. Anggota Laboratorium
d. *Lab Based Education* (LBE)

Nilai *K* untuk pengujian ini adalah 300, 350, 400, 450 dan 500, sementara *c* adalah 0,1 sampai 0,9.

Tabel 4.21 Distribusi variasi asal Jurusan anggota Peneliti.

Data	Variasi Asal Jurusan Anggota (n Jurusan)						Jumlah (kelompok)
	n=1	n=2	n=3	n=4	n=5	n=6	
Penelitian Dosen 2012	251	57	10	5	2	0	325
Penelitian Dosen 2013	293	78	22	5	0	0	398
Anggota Laboratorium	160	6	1	0	0	0	167
LBE	38	9	3	0	0	0	50

Tabel 4.22 Nilai *Entropy* Penelitian Dosen tahun 2012

Nilai c	*Entropy* Penelitian Dosen tahun 2012				
	K=300	*K*=350	*K*=400	*K*=450	*K*=500
0,1	0,374	0,371	0,375	0,378	0,378
0,2	0,379	0,381	0,370	0,376	0,381
0,3	0,387	0,391	0,380	0,384	0,384
0,4	0,385	0,388	0,381	0,390	0,388
0,5	0,386	0,393	0,385	0,383	0,387
0,6	0,390	0,395	0,384	0,384	**0,393**
0,7	0,387	0,395	0,392	0,391	0,387
0,8	0,386	**0,397**	0,389	**0,391**	0,391
0,9	**0,398**	0,385	**0,394**	0,388	0,393

Data Penelitian Dosen tahun 2012 memiliki 325 kelompok Peneliti dengan variasi asal Jurusan seperti pada Tabel 4.21, dimana kelompok yang ada didominasi kelompok yang anggotanya berasal dari Jurusan yang sama. Pada Tabel 4.22 hasil pengelompokan yang memiliki nilai *Entropy* tertinggi sebesar 0,398 adalah

pengelompokan dengan nilai *K* = 300 dan *c* = 0,9, hal ini mengindikasikan bahwa kelompok Penulis ini adalah kelompok yang paling mririp dengan data Peneliti pada Penelitian Dosen tahun 2012. Secara keseluruhan nilai *Entropy* akan naik seiring dengan kenaikan nilai *c* (Gambar B.1), hal ini menggambarkan bahwa kelompok Penulis yang dibentuk dengan memprioritaskan bobot kesamaan Penulis akan menghasilkan kelompok yang mirip dengan data Penelitian Dosen tahun 2012.

Berdasarkan data Penelitian Dosen tahun 2013, hasil pengelompokan yang memiliki nilai *Entropy* tertinggi sebesar 0,430 adalah pengelompokan dengan nilai *K* = 350 dan *c* = 0,8 (Tabel 4.23), hal ini mengindikasikan bahwa kelompok Penulis ini adalah kelompok yang paling mririp dengan data Peneliti pada Penelitian Dosen tahun 2013. Secara keseluruhan nilai *Entropy* akan naik seiring dengan kenaikan nilai *c* (Gambar B.2), dimana nilai *Entropy* dari setiap nilai *K* berada pada nilai *c* lebih besar dari atau sama degan 0,7, hal ini menggambarkan bahwa kelompok Penulis yang dibentuk dengan memprioritaskan bobot kesamaan Penulis akan menghasilkan kelompok yang mirip dengan data Penelitian Dosen tahun 2013.

Tabel 4.23 Nilai *Entropy* Penelitian Dosen tahun 2013

Nilai c	***Entropy* Penelitian Dosen tahun 2013**				
	***K*=300**	***K*=350**	***K*=400**	***K*=450**	***K*=500**
0,1	0,411	0,415	0,410	0,414	0,418
0,2	0,419	0,421	0,414	0,415	0,417
0,3	0,418	0,422	0,415	0,424	0,421
0,4	0,424	0,415	0,427	0,422	0,427
0,5	0,420	0,419	0,421	0,418	0,420
0,6	0,425	0,425	0,418	0,416	0,423
0,7	0,422	0,430	**0,427**	**0,425**	0,425
0,8	0,423	**0,430**	0,423	0,419	0,427
0,9	**0,425**	0,424	0,426	0,423	**0,428**

Data anggota Laboratorium memiliki keistimewaan yaitu asal Jurusan anggotanya adalah mayoritas dari satu Jurusan, hal ini sehubungan dengan keberadaan Laboratorium sebagai salah satu fasilitas yang dimiliki Jurusan. Hal ini yang mengakibatkan nilai *Entropy* anggota Laboratorium menjadi kecil, karena data kelompok Penulis sebagian besar berasal dari beberapa Jurusan. Pada Tabel

4.23, nilai *Entropy* tertinggi sebesar 0,271 saat nilai K = 500 dan c = 0,9. Kelompok yang terbentuk dari nilai K = 500 dan c = 0,9 (Tabel 4.24) adalah kelompok yang paling mirip diantara kelompok lainnya. Secara keseluruhan nilai *Entropy* terbaik diperoleh dari nilai c lebih besar atau sama dengan 0,8 (Gambar B.3), hal ini menunjukkan bahwa dengan memprioritaskan nilai kesamaan topik antar Penulis memberikan hasil yang menyerupai data anggota Laborarorium, karena Penulis yang menjadi anggota Laboratorium biasanya memiliki topik yang mirip dan juga kolaborasi yang tinggi, tetapi anggotanya masih terbatas pada satu Jurusan saja.

Tabel 4.24 Nilai *Entropy* Anggota Laboratorium

Nilai c	***Entropy* Anggota Laboratorium**				
	***K*=300**	***K*=350**	***K*=400**	***K*=450**	***K*=500**
0,1	0,255	0,253	0,251	0,255	0,256
0,2	0,252	0,252	0,246	0,259	0,254
0,3	0,252	0,259	0,254	0,260	0,254
0,4	0,251	0,254	0,257	0,261	0,261
0,5	0,250	0,254	0,248	0,252	0,258
0,6	0,256	0,260	0,250	0,256	0,256
0,7	0,253	0,263	0,258	0,257	0,258
0,8	0,260	**0,270**	0,263	0,256	0,262
0,9	**0,262**	0,268	**0,265**	**0,264**	**0,271**

Lab Based Education (LBE) adalah komunitas Peneliti di lingkungan ITS dimana anggotanya adalah Dosen, Mahasiswa Pascasarjana dan Pakar atau Praktisi dari luar ITS. LBE adalah salah satu bentuk kerjasama ITS Surabaya dengan Kumamoto University, dibawah naungan *Japan International Cooperation Agency* (JICA). Fokus penelitian kelompok Peneliti pada LBE adalah memberi solusi pada permasalahan non-akademis atau masalah di luar ITS Surabaya. Nilai *Entropy* untuk LBE sangat kecil jika dibandingkan dengan dala lainnya, karena Penulis (Dosen) yang menjadi anggota LBE masih sedikit dari total 751 Penulis hanya ada 173 Penulis saja. Nilai *Entropy* tertinggi sebesar 0,120 diperoleh saat K = 500 dan c = 0,2 (Tabel 4.25), tetapi perbedaan nilai *Entropy* secara keseluruhan tidak signifikan (Gambar B.4), hal ini disebabkan karena irisan antara kelompok Penulis dengan kelompok Peneliti sangat kecil antara satu dan dua anggota saja.

Tabel 4.25 Nilai *Entropy Lab Based Education*

Nilai c	***Entropy* Anggota *Lab Based Education***				
	***K*=300**	***K*=350**	***K*=400**	***K*=450**	***K*=500**
0,1	0,115	0,114	0,112	0,112	0,111
0,2	0,112	0,114	0,113	0,113	**0,120**
0,3	0,115	0,115	0,112	0,112	0,112
0,4	0,117	0,113	0,112	**0,117**	0,112
0,5	0,114	0,109	0,111	0,107	0,111
0,6	0,112	0,113	0,109	0,111	0,113
0,7	0,114	0,116	**0,113**	0,110	0,114
0,8	**0,119**	**0,118**	0,110	0,114	0,118
0,9	0,119	0,117	0,111	0,112	0,119

Secara keseluruhan, hasil pengujian menunjukkan bahwa hasil pengelompokan *Hypergraph-Partitioning* pada *Co-Authorship Graph* yang dibentuk dengan nilai *c* yang besar memberikan nilai *Entropy* yang besar pula, hal ini dapat dilihat dari nilai *Entropy* terbaik pada tiap nilai *K* pada setiap pengujian hampir selalu didapatkan dari nilai *c* yang lebih besar dari atau sama dengan 0,6. Nilai *c* yang besar ini mengindikasikan bahwa dengan memprioritaskan bobot kesamaan topik antar Penulis memberikan hasil pengelompokkan yang baik.

4.6 Kendala Uji Coba

Dalam melakukan uji coba dan validasi, terdapat beberapa kendala yang mempengaruhi hasil pengujian di atas. Beberapa kendala tersebut dapat dijabarkan sebagai berikut :

1. Pada informasi Abstrak karya tulis masih terdapat banyak kesalahan pengetikan sehingga harus dibenahi secara manual sehingga menyebabkan praproses harus dilakukan beberapa kali untuk mendapatkan hasil yang baik.
2. Banyak Dokumen untuk beberapa Penulis masih sedikit sehingga topik yang didapatkan belum dapat merepresentasikan keadaan yang sebenarnya. Setidaknya banyak Dokumen per Penulis adalah minimal 10% dari keseluruhan Dokumen.

3. Data Peneliti dari LPPM ITS Surabaya masih menggunakan nama peneliti sebagai *primary key*, dimana beberapa terdapat pula kesalahan pengetikan sehingga menyebabkan beberapa kesalahan pada saat penyusunan data Peneliti, misalnya terdapat nama Budie Santosa (Teknik Sipil), Budi Santoso (Teknik Industri), Budi Santosa (Teknik Industri) bila terjadi kesalahan pengetikan maka akan menjadi kesalahan identifikasi anggota kelompok Peneliti. Seharusnya yang dipakai sebagai *primary key* adalah Nomor Induk Dosen Nasional (NIDN) seperti yang digunakan pada informasi Dosen.

[Halaman ini sengaja dikosongkan]

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut.

1. Metode *Hypergraph-Partitioning* dapat digunakan untuk membantu mengelompokkan Penulis berdasarkan kesamaan topik antar Penulis.
2. Dari hasil uji coba 1, secara keseluruhan tidak dapat dipastikan hubungan antara nilai *c* dan *K* dengan kualitas kelompok, dimana karakteristik Dokumen yang dipakai sangat mempengaruhi.
3. Dari hasil uji coba 1 dan 2, diketahui bahwa semakin banyak topik maka nilai kesamaan topik antar Penulis juga meningkat, tetapi tidak memberikan jaminan hasil pengelompokan akan memiliki kualitas kelompok yang baik atau struktur kelompok yang baik pula.
4. Dari hasil uji coba 3, Secara keseluruhan hasil pengelompokkan yang miliki tingkat validitas tertinggi hampir selalu didapatkan dari nilai *c* yang lebih besar dari atau sama dengan 0,6, untuk setiap nilai *K*. Nilai *c* yang besar mengindikasikan bahwa dengan memprioritaskan bobot kesamaan topik antar Penulis memberikan hasil pengelompokkan yang baik.

5.2 Saran

Berdasarkan hasil penelitian ini, saran yang dapat diberikan agar diperoleh hasil yang lebih baik adalah :

1. Banyak Dokumen untuk tiap Penulis minimal 5% dari keseluruhan Dokumen.
2. Banyak Topik pada proses ekstraksi topik sesuai kebutuhan atau asumsi, karena mempengaruhi lamanya waktu ekstraksi topik.

[Halaman ini sengaja dikosongkan]

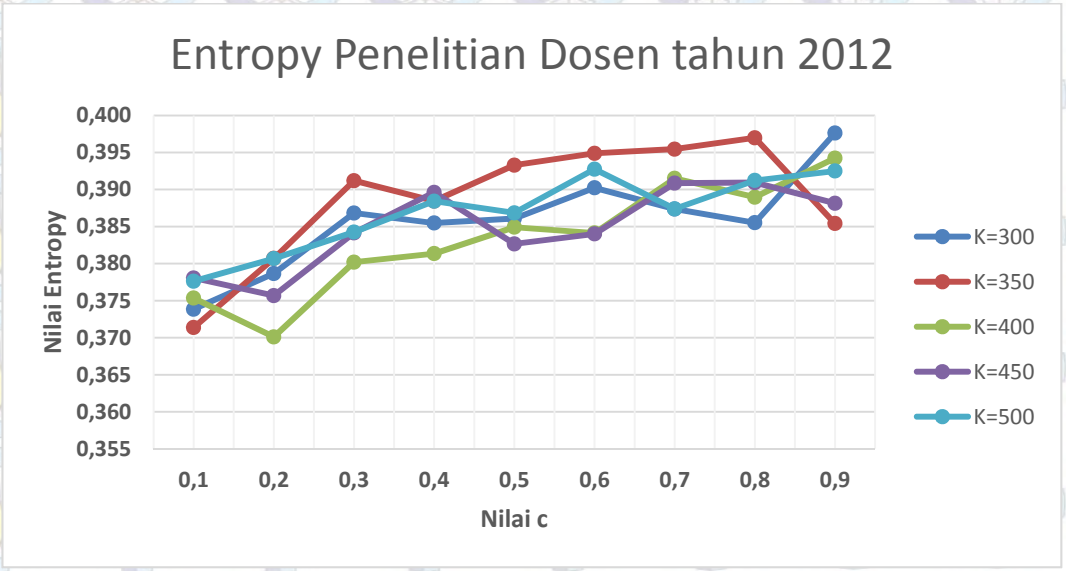
LAMPIRAN A

Tabel A.1 Dokumen dan probabilitas topik Penulis dari bu Diana Purwitasari.

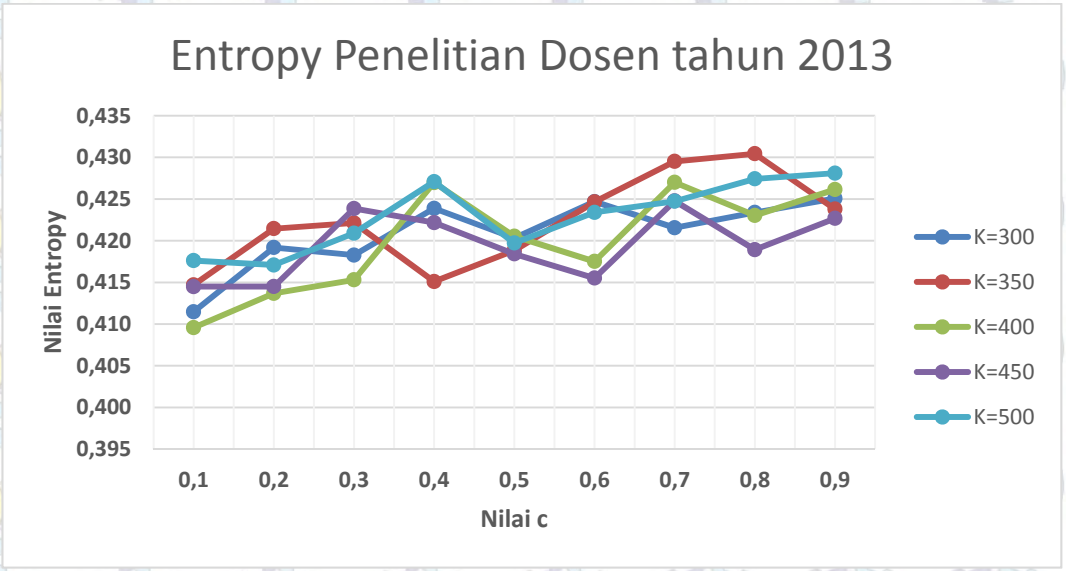
ID Dokumen	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
5105100076	8,08	4,62	11,78	8,61	4,45	10,08	3,23	2,89	2,42	**37,90**	3,10	2,86
5105201004	3,50	3,12	3,11	3,70	6,14	**50,95**	13,44	2,35	3,79	2,99	3,19	3,71
5106100022	2,08	2,35	12,05	11,61	2,41	**53,43**	2,62	2,11	2,42	4,55	2,09	2,26
5106100040	3,95	**35,61**	3,50	2,25	33,10	2,26	3,48	2,34	2,03	4,54	2,12	4,83
5106100043	3,68	22,96	2,42	5,99	9,83	6,70	2,83	2,32	2,42	2,80	2,34	**35,70**
5106100048	3,98	4,11	14,82	5,32	4,23	**41,01**	4,61	4,75	4,59	3,98	3,98	4,62
5106100059	2,96	2,20	5,46	4,11	2,23	9,81	2,16	6,26	7,77	8,39	2,69	**45,96**
5106100075	2,70	13,83	**41,34**	8,56	3,25	5,16	6,35	3,95	4,59	3,39	2,94	3,96
5106100090	2,92	**35,39**	2,75	3,99	21,63	4,86	3,68	2,69	2,56	2,67	2,57	14,30
5106100095	9,31	2,22	21,03	13,18	3,90	**30,72**	3,99	2,76	2,72	4,92	2,51	2,75
5106100099	3,62	5,10	6,04	3,59	3,89	**37,80**	3,23	5,10	15,02	9,15	4,23	3,24
5106100108	3,22	12,76	3,03	**29,78**	14,40	4,80	15,25	3,10	3,14	3,34	2,84	4,34
5106100151	2,71	3,89	3,85	9,23	5,67	**52,93**	8,86	1,99	3,08	2,20	3,12	2,47
5106100158	3,43	8,32	3,23	12,54	7,60	**39,53**	6,40	3,33	4,34	3,24	4,16	3,88
5106100801	3,50	4,22	3,43	13,90	3,28	**47,78**	5,41	3,38	3,68	3,94	3,71	3,77
5106100802	3,05	3,61	3,45	4,60	13,11	24,26	**29,99**	4,02	4,05	3,86	2,97	3,03
5106100803	5,06	7,19	3,12	7,50	6,04	**53,42**	2,84	3,12	3,39	2,80	2,69	2,83
5107100001	1,88	3,03	2,24	7,81	4,76	**41,31**	15,32	2,52	1,94	11,81	2,13	5,26
5107100005	3,32	2,47	14,15	7,24	3,57	**40,50**	9,39	2,52	3,03	3,91	2,65	7,23
5107100012	5,42	2,88	**32,88**	12,74	8,27	5,53	16,77	2,80	2,49	3,02	4,25	2,95
5107100057	2,60	4,14	4,14	6,16	3,20	**62,12**	3,10	2,50	3,11	3,27	2,46	3,19
5107100069	3,55	3,69	16,04	6,80	5,78	**40,22**	4,55	3,27	3,65	4,45	4,36	3,65
5107100090	2,12	2,26	3,28	6,11	16,79	**53,07**	2,41	2,46	3,06	3,05	2,37	3,02
5107100092	3,01	2,74	4,01	**26,12**	8,33	16,18	5,45	4,86	2,78	19,33	3,08	4,11
5107100134	3,99	8,37	9,03	3,35	2,87	**41,62**	2,65	3,88	5,72	4,80	4,36	9,36
5107100701	3,88	4,55	5,19	5,62	8,99	**35,57**	19,13	3,42	3,84	3,40	2,86	3,56
5108100025	3,95	2,21	18,99	5,09	4,48	10,85	**31,25**	2,31	2,16	6,45	8,76	3,50

ID Dokumen	Topik-0	Topik-1	Topik-2	Topik-3	Topik-4	Topik-5	Topik-6	Topik-7	Topik-8	Topik-9	Topik-10	Topik-11
5108100077	21,53	1,92	10,27	10,29	1,88	**42,06**	1,91	2,59	1,95	1,84	1,82	1,93
5108100136	3,38	1,78	**62,81**	2,81	4,12	8,18	5,31	1,92	1,97	3,68	1,83	2,20
5108100142	5,93	2,84	15,65	8,46	3,41	**37,40**	4,84	3,37	4,30	5,78	2,83	5,19
5108100196	4,49	2,95	4,22	**42,30**	6,04	13,95	6,57	3,02	5,73	4,68	2,89	3,15
5108100503	2,66	3,06	2,43	2,40	**67,83**	2,93	3,03	2,41	2,67	3,07	3,07	4,44
5108100510	3,75	13,36	5,08	4,39	**40,03**	5,38	5,90	3,83	4,04	4,26	3,82	6,16
5108100514	3,10	21,91	7,57	4,57	**32,64**	7,45	3,39	4,61	3,23	3,81	2,86	4,87
5108100601	7,92	4,30	**35,62**	7,67	4,14	5,11	10,02	5,55	4,39	4,89	4,40	5,99
5108100606	18,42	2,73	2,87	3,02	**44,95**	2,87	4,58	3,17	2,83	3,01	3,16	8,39
5108100607	6,02	9,88	4,91	3,78	6,68	4,59	7,01	3,50	3,56	3,69	3,39	**43,00**
5108201016	7,58	15,68	2,22	14,24	**33,93**	2,98	9,16	2,46	2,36	2,46	4,04	2,89
5108201030	3,02	**49,50**	2,08	2,65	21,37	2,51	7,00	2,14	2,23	2,49	2,45	2,56
5109201013	3,21	2,16	1,95	2,40	3,09	**73,39**	2,53	2,16	2,17	2,77	2,01	2,16
5109201025	3,65	2,54	9,99	7,49	6,31	**47,10**	11,49	2,33	2,02	2,92	2,06	2,11
5110201003	4,88	4,60	3,78	6,32	**26,17**	6,52	19,05	3,37	3,45	3,84	10,44	7,57

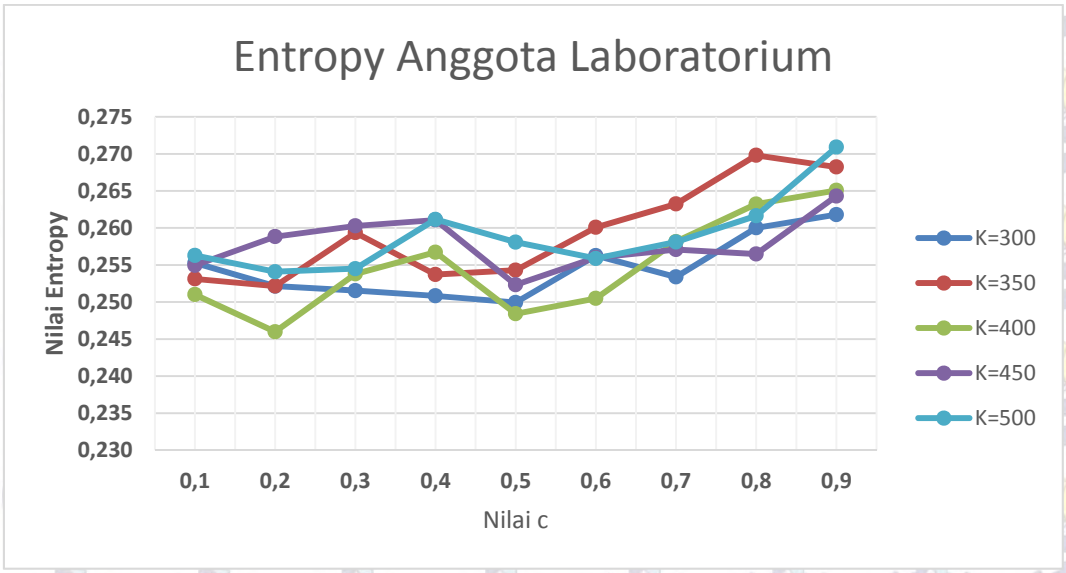
LAMPIRAN B



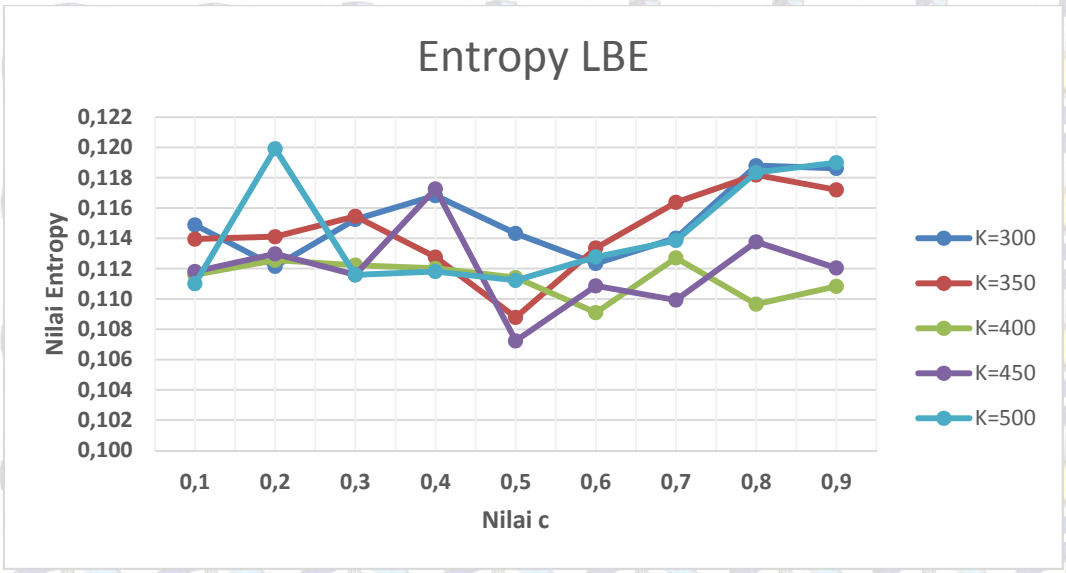
Gambar B.1 Grafik nilai *Entropy* data Penelitian Dosen tahun 2012



Gambar B.2 Grafik nilai *Entropy* data Penelitian Dosen tahun 2013



Gambar B.3 Grafik nilai *Entropy* data Anggota Laboratorium



Gambar B.4 Grafik nilai *Entropy* data *Lab Based Education* (LBE)

DAFTAR PUSTAKA

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).

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.

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

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

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.

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

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.

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).

Indra L, Daniel S, Arrie *K* and Agus ZA. “Multidocument Summarization Based on Sentence *Cluster*ing Improved Using Topic Words”, *Jurnal Ilmiah Teknologi Informasi*. Vol. 12, No. 2, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember, Agustus 2014.

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.

Papa, David A., and Igor L. Markov. "*Hypergraph Partitioning* and *cluster*ing."*Approximation algorithms and metaheuristics* (2007): 61-1.

Tala, Fadillah Z., A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Institute for Logic, *Language and Computation Universeit Van Amsterdam.* (2003)

Vivit W.R. "Kolaborasi dan *Graph* 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

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

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

Zhao, Y., & Karypis, G.. *Criterion functions for document clustering: Experiments and analysis* (pp. 01-40). Technical report. (2001)