BAB III

METODOLOGI PENELITIAN

1.1 Desain Penelitian

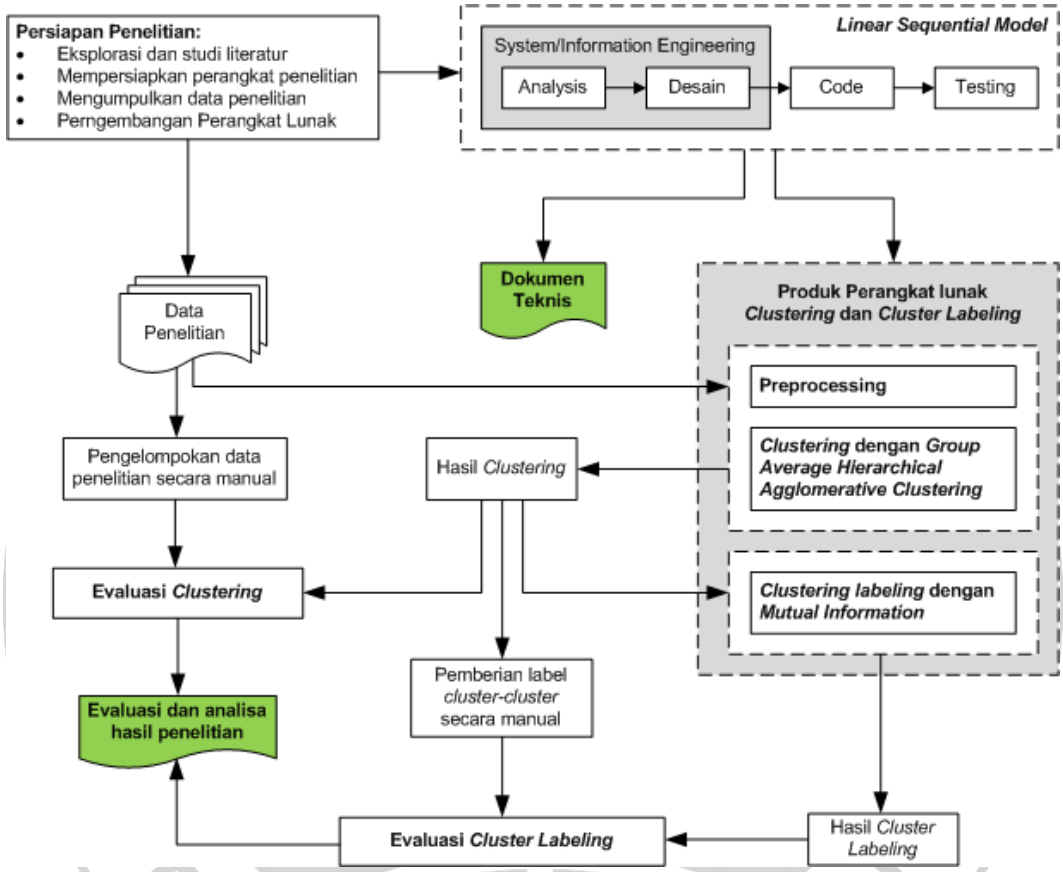
Desain penelitian adalah tahapan atau gambaran yang akan dilakukan dalam penelitian. Desain penelitian dibuat untuk memudahkan pelaksanaan tahap-tahap penelitian. Desain penelitian pada penelitian ini adalah:

1.1.1 Persiapan Penelitian

1. Eksplorasi dan studi literatur-literatur yang berhubungan dengan pendeteksian topik berita pada aliran berita online
2. Mempersiapkan perangkat penelitian berupa lingkungan penelitian, komputer, perangkat lunak pendukung, perangkat-perangkat bahasa pemrograman
3. Mengumpulkan data penelitian.
4. Pengembangan perangkat lunak dengan Linear Sequential Model yang terdiri dari langkah-langkah sebagai berikut:
    a. Analisis
    b. Perancangan
    c. Implementasi
    d. Pengujian

Hasil dari langkah ini adalah data penelitian, perangkat lunak dan dokumen teknis perangkat lunak.

Gambar 3.1 berikut adalah digram seluruh proses penelitian:



Gambar 3.1 *Desain Penelitian*

1.1.2 Implementasi Penelitian

Penelitan Clustering

Pada penelitian *clustering* ini, data penelitian yang dibutuhkan adalah dokumen-dokumen berita dan kelompok-kelompok dari dokumen-dokumen berita berita tersebut yang dikelompokkan secara manual. Algoritma yang digunakan dalam penelitian *clustering* ini adalah algoritma GA-HAC. Setelah proses

clustering dilakukan, seluruh *cluster* yang dihasilkan kemudian dibandingkan dengan pengelompokan manual untuk evaluasi.

Penelitian Cluster Labeling

Pada penelitian *cluster labeling* ini, data penelitian yang dibutuhkan adalah seluruh *cluster* hasil proses *clustering* secara otomatis dengan GA-HAC dan label dari tiap *cluster* yang ditetapkan secara manual. Algoritma yang digunakan dalam penelitian *clustering* ini adalah algoritma Mutual Information. Setelah proses *cluster labeling*, evaluasi dilakukan dengan membandingkan label-label yang dihasilkan oleh algoritma Mutual Information secara otomatis dengan label-label yang ditentukan secara manual oleh manusia.

1.2 Alat dan Bahan Penelitian

Berdasarkan kebutuhan-kebutuhan diatas, maka ditentukan bahwa alat dan bahan yang digunakan pada penelitian ini adalah sebagai berikut:

1.2.1 Perangkat Lunak

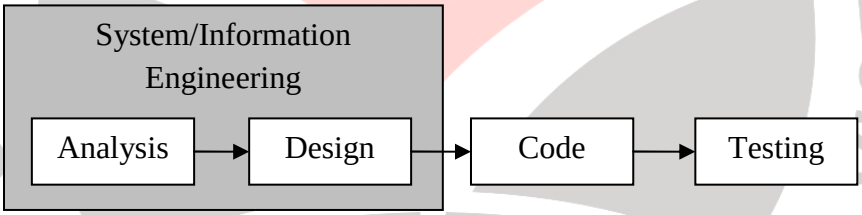
1. Sistem Operasi Ubuntu 10.10
2. JDK *(Java Development Kit)*
3. JRE (*Java Runtime Environment*)
4. Lucene 2.0
5. iPOSTagger (v. 1.0)
6. XHTML Tree Path (v. Alpha)
7. Nutch 1.2

1.2.2 Perangkat Keras

1. Processor Intel® Core™ 2 Duo T6600
2. RAM 4 GB
3. Hard Disk 5 GB

1.3 Metode Pengembangan Perangkat Lunak

Prosedur pengembangan perangkat lunak yang digunakan pada penelitian ini adalah menggunakan metode *linear sequential model*. *Linear sequential model* atau *waterfall model* merupakan metode pengembangan perangkat lunak dengan pendekatan sistematis dan berurutan. Prosedur pengembangan perangkat lunak ini digambarkan sebagai Gambar 3.2 berikut:



Gambar 3.2 *Metode Pengembangan Perangkat Lunak Linear Sequential Model*

1.3.1 *System/Information Engineering*

Perangkat lunak selalu menjadi bagian dari sistem (atau bisnis) yang lebih besar, karenanya, pengembangan perangkat lunak dengan metode *linear sequential model* dimulai dengan membuat suatu daftar kebutuhan dari seluruh elemen sistem kemudian mengalokasikan beberapa sub-bagian dari kebutuhan tersebut kepada perangkat lunak yang dikembangkan. *System/Information engineering* terdiri dari dua sub-bagian, yaitu

Software requirements analysis

Software requirement analysis adalah analisa daftar kebutuhan perangkat lunak yang terdiri dari:

1. Analisa ranah perangkat lunak
2. Perkiraan fungsi-fungsi yang dibutuhkan
3. Perilaku perangkat lunak yang diharapkan
4. Performa perangkat lunak yang dibutuhkan
5. Kebutuhan antarmuka perangkat lunak

Daftar kebutuhan dari sistem dan perangkat lunak ini didokumentasikan dan ditinjau oleh calon pengguna.

Design

Design perangkat lunak umumnya adalah sebuah proses multi langkah yang memfokuskan pada empat atribut dari sebuah program:

1. Struktur dan aliran data
2. Arsitektur perangkat lunak
3. Representasi antarmuka
4. Detail algoritma secara prosedural

Proses desain ini menterjemahkan daftar kebutuhan (*requirements*) kedalam representasi rancangan perangkat lunak yang dapat di revisi, ini dilakukan untuk memastikan kualitas sebelum tahap implementasi dimulai. Seperti *Software requirement analysis*, design didokumentasikan dan dijadikan bagian dari konfigurasi perangkat lunak.

1.3.2 *Code Implementation*

Code Implementation adalah implementasi dari hasil proses desain kedalam bentuk kode yang dapat dibaca oleh mesin. Dalam proses implementasi ini, hasil proses desain disesuaikan dengan sifat, perilaku dan kebutuhan dari bahasa pemrograman yang digunakan. Hasil dari proses code implementation ini adalah perangkat lunak yang dapat dioperasikan.

1.3.3 *Testing*

Setelah kode telah diimplementasikan, perangkat lunak yang dapat dioperasikan di uji coba. Proses uji coba ini memfokuskan pada logika internal dan fungsi eksternal perangkat lunak. uji coba logika internal dilakukan untuk memastikan bahwa seluruh pernyataan yang diimplementasikan telah diuji dan fungsi eksternal dilakukan untuk memastikan bahwa dengan input yang telah didefinisikan, perangkat lunak dapat memberikan hasil sesuai dengan hasil yang didefinisikan pada daftar kebutuhan.

1.4 Prosedur Pengerjaan Penelitian

1.4.1 Persiapan Penelitian

Eksplorasi dan Studi Literatur

Penelitian ini dimulai dengan pengumpulan pustaka penelitian yang membahas mengenai pendeteksian topik pada kumpulan dokumen-dokumen. Pada penelusuran tahap awal, ditemukan bahwa penelitian tentang *Topic Detection and*

Tracking (TDT) adalah penelitian yang paling berkesesuaian secara parsial dengan topik penelitian yang dipilih.

Pada penelitian tentang TDT, teknik umum yang digunakan untuk menentukan topik dokumen-dokumen berita adalah teknik *clustering*, namun teknik *clustering* tersebut tidak sepenuhnya memenuhi ekspektasi penelitian, yaitu cara merepresentasikan dokumen-dokumen berita agar dapat dimengerti oleh pembaca tanpa harus membaca seluruh dokumen pada seluruh *cluster* hasil *clustering*. Oleh karena itu, penelusuran lanjutan dilakukan dengan mencari teknik representasi yang dapat digabungkan dengan teknik *clustering* dan ditemukan bahwa teknik yang berkesesuaian dengan hal tersebut adalah teknik *cluster labeling*.

Setelah itu, penelusuran lanjutan dilakukan dengan mengeksplorasi penelitian-penelitian yang berhubungan dan mempelajari dasar-dasar teori dari penelitian-penelitian tersebut, ini termasuk berbagai variasi implementasi teknik yang didapatkan dari referensi pustaka penelitian dan pencarian di internet.

Persiapan Perangkat Penelitian

Setelah penelusuran dan pengumpulan pustaka penelitian, pengerjaan penelitian dilanjutkan dengan mempersiapkan alat dan bahan penelitian berupa lingkungan operasi, berbagai perangkat keras dan perangkat lunak yang kira-kira dapat memenuhi kebutuhan operasional dan pelaksanaan penelitian. Pada tahap ini ditentukan bahwa perangkat keras dan lingkungan operasional adalah komputer desktop umum yang dengan sistem operasi Ubuntu Linux.

Perangkat lunak yang digunakan pada penelitian ini dipilih berdasarkan kegunaan perangkat lunak tersebut dalam memenuhi kebutuhan penelitian. Beberapa kebutuhan dari penelitian ini adalah:

1. Kebutuhan Pengumpulan Data Penelitian
    a. Kemampuan penyimpanan dan pencarian data teks
    b. Kemampuan pengambilan data-data aliran berita secara otomatis
    c. Kemampuan pengolahan dokumen online (HTML)
2. Kebutuhan Pemrosesan
    a. Kemampuan penyimpanan dan pencarian data teks
    b. Kemampuan pemrosesan bahasa alami (*Natural Languange Processing*) untuk bahasa Indonesia.

Mengumpulkan Data Penelitian

Sumber data yang digunakan pada penelitian ini adalah aliran berita yang didapatkan selama 48 jam dari tanggal 25-04-2011 jam 14:00 sampai dengan tanggal 27-04-2011 jam 13:00 dari situs-situs berita:

- http://nasional.kompas.com/
- http://nasional.vivanews.com
- http://www.antaranews.com/berita
- http://www.republika.co.id/berita/nasional

Data penelitian ini didapatkan dengan mengimplementasikan metode *time window* dengan waktu pengambilan tiap interval 1 jam dan luas waktu window 24 jam. Penentuan ini mengakibatkan dokumen-dokumen berita yang melebihi waktu

24 jam dari pengambilan terakhir tidak diikut sertakan dalam koleksi dokumen terbaru. Untuk menghindari duplikasi dokumen berita, digunakan acuan berdasarkan URL artikel berita, dimana artikel berita yang memiliki URL sama dianggap sebagai dokumen yang sama.

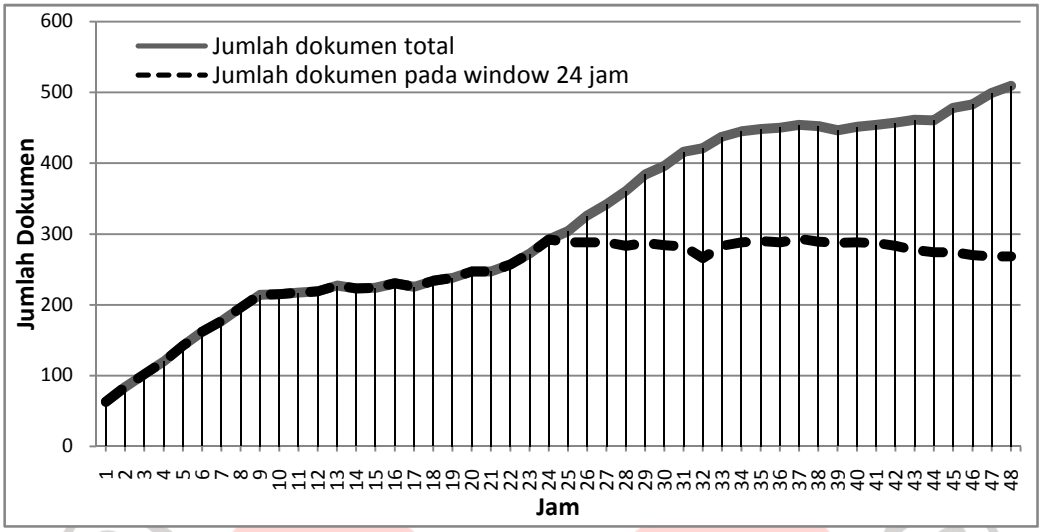
Informasi-informasi yang digunakan pada penelitian ini adalah judul berita dan isi berita saja, informasi-informasi lain seperti ilustrasi gambar, tautan berita terkait, kata kunci dan lain-lain tidak digunakan.

Hasil dari implementasi metode *time window* dalam mengumpulkan data penelitian selama 48 jam dituliskan pada Tabel 3.1 berikut:

Tabel 3.1 *Jumlah dokumen pada tiap jam pengambilan*

Jam	Jumlah dokumen pada window 24 jam	Jumlah Dokumen total	Jam	Jumlah dokumen pada window 24 jam	Jumlah dokumen total
1	63	63	25	288	304
2	84	84	26	288	326
3	102	102	27	288	342
4	120	120	28	283	361
5	142	142	29	287	384
6	162	162	30	284	396
7	177	177	31	282	416
8	196	196	32	266	421
9	214	214	33	283	437
10	215	215	34	288	445
11	217	217	35	290	448
12	219	219	**36**	**288**	**450**
13	227	227	37	293	454
14	223	223	38	289	452
15	224	224	39	287	446
16	230	230	40	288	451
17	225	225	41	287	454
18	234	234	42	283	457
19	238	238	43	277	461
20	247	247	44	274	460
21	247	247	45	274	478
22	257	257	46	270	483
23	272	272	47	268	499
24	**292**	**292**	**48**	**268**	**509**

Gambar 3.3 berikut adalah grafik jumlah dokumen untuk tiap jam pengambilan:



Gambar 3.3 *Grafik Jumlah dokumen pada tiap jam pengambilan*

Pengembangan Perangkat Lunak

Pengembangan perangkat lunak ini dilakukan pada lingkungan bahasa pemrograman Java. Tahap ini melibatkan beberapa perangkat bahasa pemrograman yaitu pustaka-pustaka standar bahasa pemrograman Java dan pustaka pemrosesan *Part-of-Speech Tagger* untuk Bahasa Indonesia. Pada tahap ini pula antar muka perangkat lunak dibuat, salah satu kebutuhan dari perangkat lunak ini adalah kebebasan memilih dokumen-dokumen yang akan diproses.

1.4.2 Implementasi Penelitian

Penelitan Clustering

Pada penelitian *clustering* ini, data penelitian yang dibutuhkan adalah dokumen-dokumen berita dan kelompok-kelompok dari dokumen-dokumen berita

berita tersebut yang dikelompokkan secara manual. Proses yang terlibat dalam analisa hasil penelitian *clustering* ini adalah:

a. Tahap *preprosessing* yang terdiri dari *tokenization* untuk mendapatkan tiap *term* isi dokumen berita.
b. Tahap *clustering* yang terdiri dari representasi dokumen berita sebagai vektor, penghitungan nilai TF-IDF, penghitungan *cosine similarity* dan terakhir proses *clustering* dengan algoritma GA-HAC dengan *narutal clustering*.

Setelah proses *clustering* dilakukan, seluruh *cluster* yang dihasilkan kemudian dibandingkan dengan pengelompokan manual untuk evaluasi. Hasil perbandingan tersebut menggunakan metrik-metrik evaluasi *Purity*, *Precision*, *Recall* dan F_1 *Measure*.

Penelitian Cluster Labeling

Pada penelitian *cluster labeling* ini, data penelitian yang dibutuhkan adalah seluruh *cluster* hasil proses *clustering* secara otomatis dengan GA-HAC dan label dari *cluster* yang ditetapkan secara manual. Proses yang terlibat dalam analisa hasil penelitian *cluster labeling* ini adalah:

a. *Phrases extractor* yang didalamnya melibatkan POS Tagger untuk mendapatkan frasa-frasa calon *cluster label*.
b. Mutual Information menggunakan Laplace Correction untuk menentukan *cluster label* yang paling sesuai dengan tiap *cluster*.

Setelah proses *cluster labeling*, evaluasi dilakukan dengan membandingkan label-label yang dihasilkan oleh algoritma Mutual Information secara otomatis dengan label-label yang ditentukan secara manual oleh manusia. Hasil perbandingan tersebut menggunakan metrik-metrik evaluasi *Match@N*, *P@N*, *MRR* dan *MTRR*.