BAB 2 LANDASAN TEORI

Bab ini berisi landasan teori tentang sistem perolehan informasi, sistem perolehan informasi XML, undang-undang Republik Indonesia berformat XML, *open source search engine*, metode evaluasi, dan pembahasan.

2.1 Sistem Perolehan Informasi

Pengertian dari sistem perolehan informasi sangat beragam dan luas. Sebuah aksi sederhana seperti membaca daftar isi suatu buku untuk mengetahui nomor halaman bab tertentu sudah dapat dimasukkan dalam pengertian perolehan informasi. Dalam dunia akademik istilah perolehan informasi dapat diartikan sebagai sebuah upaya menemukan suatu material (biasanya dokumen) yang memenuhi kebutuhan akan informasi di antara sejumlah besar koleksi [MAN08]. Materi atau obyek dari perolehan informasi saat ini cukup beragam, seperti dokumen teks biasa (txt), dokumen dengan format html, xml, pdf, audio, video, atau gambar tidak bergerak.

Menemukan kembali suatu materi dalam koleksi yang relevan dengan kebutuhan informasi merupakan inti dari sistem perolehan informasi. Untuk itu, sebuah sistem perolehan informasi setidaknya memiliki tiga area utama, yakni pengindeksan (*indexing*), pencarian (*searching*), dan pemeringkatan (*ranking*) [MID07].

2.1.1 Pengindeksan

Indeks merupakan suatu struktur data yang merepresentasikan dan mengorganisasikan isi dari dokumen-dokumen dalam koleksi. Tujuan dari adanya indeks ini adalah untuk efisiensi atau percepatan dalam hal pencarian informasi yang ada di koleksi [MID07]. Adanya indeks ini mengurangi jumlah proses pencocokan kueri dengan dokumen koleksi. Di sinilah peran penting dari struktur data indeks. Jenis struktur data yang paling sering digunakan dalam perolehan

informasi teks adalah *inverted index*. *Inverted index* terdiri dari *vocabulary* dan *posting list*. *Vocabulary* merupakan daftar seluruh kata yang ada dalam koleksi, sementara *posting list* merupakan daftar dokumen atau posisi yang memuat setiap kata dalam *vocabulary* [MID07].

Gambar 2.1. Contoh *inverted index.*

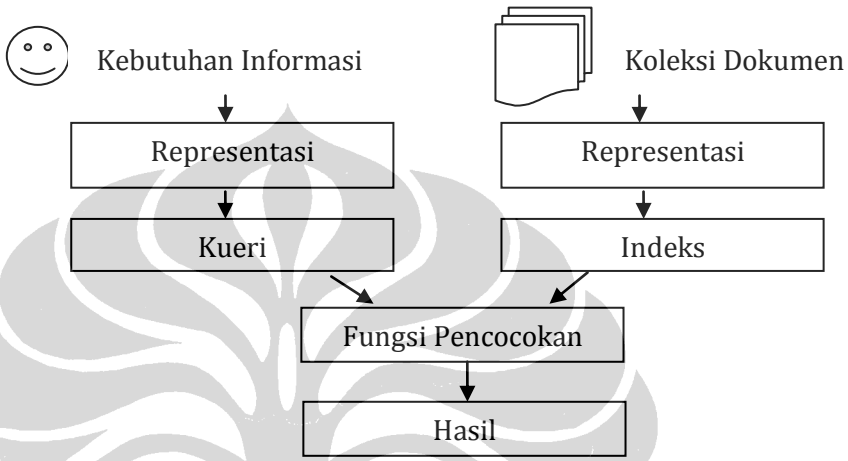
Ada beberapa fitur tambahan yang dapat dipergunakan saat proses pengindeksan, antara lain pembuangan *stopwords* dan *stemming*. Fitur pertama, pembuangan *stopwords* berfungsi mengurangi kata-kata yang sering muncul tetapi tidak memiliki arti sehingga kurang bermanfaat dalam proses perolehan informasi. Contoh dari *stopwords* antara lain "yang", "di", "ke", "dari", dan "dan". Tidak diindeksnya *stopwords* dapat membuat indeks menjadi lebih ramping.

Stemming merupakan proses yang memetakan kata-kata dengan variasi morfologi ke bentuk dasarnya. *Stemming* menghilangkan imbuhan-imbuhan yang melekat pada suatu kata. Contohnya kata "memakan", "dimakan", "termakan", dan "makanan" dirubah ke bentuk dasarnya menjadi "makan". *Stemmer* yang dikenal untuk bahasa inggris antara lain Porter Stemmer, Lovins Stemmer, dan Krovetz Stemmer. Sementara itu, *stemmer* untuk bahasa Indonesia ada algoritma stemming yang dikembangkan oleh Nazief dan Adriani [NAZ96]. Akurasi *stemmer* untuk bahasa Indonesia tersebut adalah 93%. Penggunaan *stemmer* dapat meningkatkan performa perolehan informasi dan mengurangi ukuran indeks.

2.1.2 Pencarian dan pemeringkatan

Pencarian merupakan usaha mengekstrak informasi dari indeks berdasarkan kueri dari pengguna [MID07]. Dalam proses ekstraksi tersebut, diukur seberapa besar

kemiripan antara kueri dengan dokumen dalam indeks. Untuk melakukan hal tersebut, digunakanlah model perolehan informasi. Model ini merupakan semacam pendekatan dari proses yang sesungguhnya terjadi. Model perolehan informasi mempunyai tiga komponen yakni (1) representasi dokumen; (2) representasi kueri; (3) Fungsi pencocokan antara representasi kueri dengan representasi dokumen.



Gambar 2.2. Contoh proses perolehan informasi.

Beberapa model yang dikenal dalam *information retrieval* antara lain *boolean model, vector space model (VSM), statistical language model (LSM), latent semantic analysis (LSA),* dan *inference network.* Berdasarkan model-model tersebut dikenal pula fungsi turunannya seperti TF*IDF, dan *cosine similarity model* yang berdasar pada VSM, Okapi dan KL-divergence yang merupakan turuan dari LSM, dan *Indri model* yang menggunakan gabungan antara *inference network* dan SLM [DIA06]. Keluaran dari fungsi setiap model berupa himpunan dokumen yang relevan. Pada beberapa model, keluarannya berupa daftar dokumen yang telah diberi peringkat berdasarkan nilai kecocokannya. Pemeringkatan ini sangat bermanfaat apabila jumlah dokumen hasil pencarian sangat banyak, sementara pengguna menginginkan hanya dokumen yang paling relevan.

Pada pembahasan sebelumnya, disebutkan istilah kueri. Kueri merupakan potongan informasi yang mewakili kebutuhan pengguna akan informasi yang ada di indeks. Bentuk kueri bermacam-macam, biasanya mengikuti bentuk atau

format dari koleksi. Jika koleksi berupa dokumen teks, maka kuerinya juga berupa teks. Jika koleksinya dokumen citra, musik, atau video, bisa jadi kuerinya berformat sama dengan format koleksi, tetapi yang umum digunakan adalah kueri teks.

Kueri berformat teks mempunyai variasi penulisan seperti boolean dan *proximity*. Kueri berjenis boolean adalah kueri yang disisipi dengan operator boolean seperti "AND", "OR", dan "NOT". Sementara itu, *proximity query* menginginkan kata-kata muncul dengan jarak tertentu, tetapi dengan memperhatikan urutan kata. Operator yang digunakan dalam kueri jenis ini antara lain "WITH" dan "NEAR n".

2.2 Sistem Perolehan Informasi XML

Salah satu jenis sistem perolehan informasi adalah sistem perolehan informasi XML. Sesuai dengan namanya, obyek sistem ini adalah dokumen XML. Hal yang melatarbelakangi munculnya sistem perolehan informasi XML adalah makin meluasnya penggunaan XML sebagai standar dokumen teks berstruktur. Subbab 2.2.1 membahas secara sekilas format XML, Subbab 2.2.1 membahas tantangan yang ada dalam perolehan informasi XML.

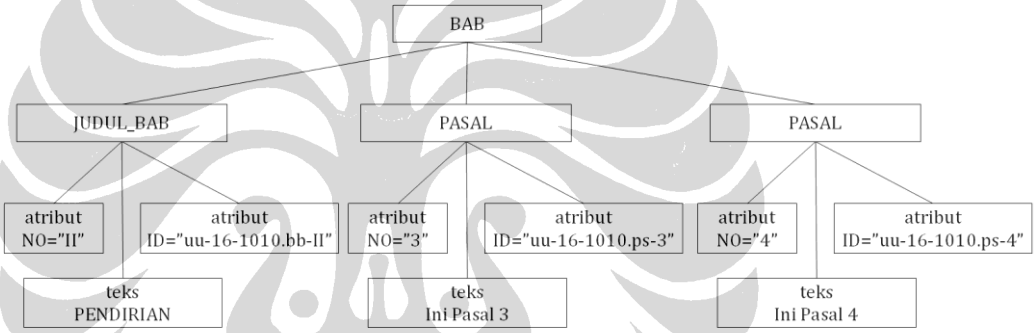
2.2.1 Format XML

Harold dan Means menyebutkan bahwa yang dimaksud *eXtensible Markup Language* (XML) adalah sebuah sintaks generik yang digunakan oleh manusia untuk menandai data dengan *tag-tag* sederhana yang dapat dibaca oleh manusia [HAR01]. Tujuannya adalah menyediakan format standar untuk dokumen-dokumen dijital. Sementara itu, Manning mendefinisikan XML sebagai dokumen yang memiliki label terurut dan dapat direpresentasikan sebagai sebuah *tree* [MAN08]. Setiap *node* dari *tree* tersebut dinamakan elemen XML yang mempunyai *tag* pembuka dan *tag* penutup. Elemen XML juga dapat memiliki atribut yang didefinisikan pada *tag* pembuka.

```
<BAB NO="II" ID="uu-16-2010.bb-II">
<JUDUL_BAB>PENDIRIAN</JUDUL_BAB>
  <PASAL NO="3" ID="uu-16-2010.ps-3">Ini Pasal 3 </PASAL>
  <PASAL NO="4" ID="uu-16-2010.ps-4">Ini Pasal 4 </PASAL>
</BAB>
```

Gambar 2.3. Contoh elemen bab

Gambar 2.3 menunjukkan sebuah elemen XML dengan *tag* pembuka <BAB> dan *tag* penutup </BAB>. Elemen tersebut mempunyai dua buah atribut, yakni NO yang bernilai "II" dan ID yang mempunyai nilai "uu-16-2010.bbII". Selain itu, elemen bab juga mempunyai dua jenis elemen anak, yakni judul bab dan pasal. Apabila elemen bab tersebut direpresentasikan dalam bentuk *tree*, maka tampilannya seperti ditunjukkan oleh Gambar 2.3.



Gambar 2.4. Representasi elemen bab pada Gambar 2.2 dalam struktur *tree.*

Gambar 2.4 menunjukkan bahwa *leaves node* dari *tree* tersebut terdiri dari teks-teks, yakni "PENDIRIAN", "Ini Pasal 3", dan "Ini Pasal 4". Tidak seperti HTML dimana pengguna hanya dapat menggunakan *tag-tag* yang sudah ada, penamaan *tag* XML dapat didefinisikan secara bebas oleh pengguna. Namun demikian, tata cara penulisan dokumen XML harus memperhatikan aturan-aturan tertentu seperti penulisan *tag* dan elemen yang tidak menyimpang dari DTD, urutan penulisan *tag*, komentar, dan atribut dengan tujuan penulisan *tag-tag* yang dibuat tertata benar dan mengikuti standar, *well-formed*. *Document Type Definition* (DTD) sendiri merupakan suatu dokumen berisi sintaks-sintaks yang menjelaskan elemen, entitas, dan atribut yang digunakan dalam dokumen XML. Aturan-aturan tersebut serta DTD dibahas secara rinci oleh Harold dan Means [HAR01].

2.2.2 Tantangan

Sistem perolehan informasi XML memungkinkan eksploitasi struktur yang ada dalam dokumen karena setiap struktur sudah diberi penanda berupa *tag XML*. Eksploitasi struktur XML membuat pengguna memperoleh bagian dokumen XML yang paling spesifik, bukan keseluruhan dokumen seperti pada *unstructured retrieval* [MAN08]. Hasil pencarian sistem ini dapat muncul dari elemen manapun dari dokumen-dokumen XML dalam koleksi.

```
<BAB NO="III" ID="uu-11-2008.bb-III">
<JUDUL_BAB>INFORMASI, DOKUMEN, DAN TANDA TANGAN
ELEKTRONIK</JUDUL_BAB>
<PASAL NO="5" ID="uu-11-2008.ps-5">
<AYAT NO="1" ID="uu-11-2008.ps-5.ay-1">
Informasi Elektronik dan/atau Dokumen Elektronik dan/atau hasil cetaknya
merupakan alat bukti hukum yang sah
</AYAT>
<AYAT NO="2" ID="uu-11-2008.ps-5.ay-2">
Informasi Elektronik dan/atau Dokumen Elektronik dan/atau hasil cetaknya
sebagaimana dimaksud pada <REF ID="uu-11-2008.ps-5.ay-1"/> merupakan
perluasan dari alat bukti yang sah sesuai dengan Hukum Acara yang berlaku di
Indonesia
</AYAT>
```

Gambar 2.5. Salah satu bab dalam UU No 11 Tahun 2008.

Gambar 2.5 merupakan potongan dari undang-undang nomor 11 tahun 2008 tentang Informasi Elektronik. Saat pengguna memberikan kueri "informasi elektronik", ada empat elemen yang dapat dikembalikan sistem, yakni undang-undang, bab, pasal atau ayat karena pada keempat elemen tersebut termuat kata-kata yang ada dalam kueri. Mengembalikan keempat elemen tersebut secara bersamaan menimbulkan redundansi hasil pencarian. Hasil yang diinginkan pengguna tentu salah satu dari keempat elemen tersebut. Menentukan elemen mana yang diberikan kepada pengguna merupakan hal yang sulit [MAN08].

Isu lain dalam perolehan informasi XML menentukan unit pengindeksan. Pemilihan unit pengindeksan memiliki pengaruh terhadap unit perolehan informasi, unit yang diperoleh pengguna dari sistem. Hanya unit yang telah diindeks saja yang dapat menjadi *retrievable unit*. Manning membahas beberapa pendekatan yang dapat digunakan dalam masalah ini, sebagai berikut.

- Mengelompokkan elemen ke dalam unit-unit pengindeksan yang tidak saling *overlap.*
- Mengindeks elemen terbesar.
- Mengindeks elemen *leaves.*
- Mengindeks semua elemen.
- Mengindeks elemen yang dianggap berharga saja.

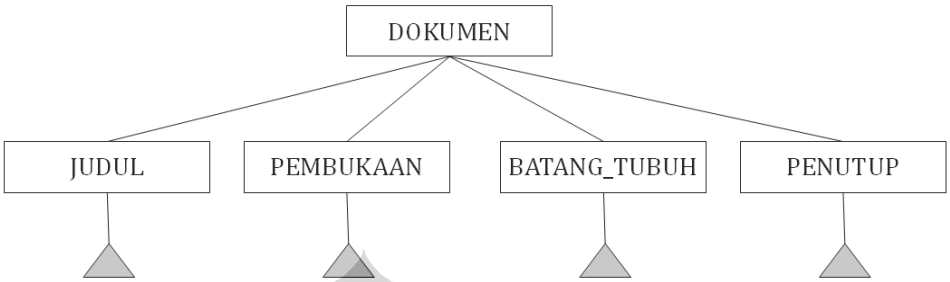
Pendekatan-pendekatan tersebut merangkum hasil beberapa penelitian yang pernah ada sebelumnya. Di antaranya adalah penelitian yang dilakukan oleh Kamps dan penelitian Sigurbjornsson. Pendekatan yang diambil dalam kedua penelitian tersebut memiliki kesamaan, yakni pengindeksan terhadap seluruh dokumen dan pengindeksan terhadap elemen dalam dokumen XML. Tujuan dari penelitian tersebut adalah mengetahui unit pengindeksan yang paling baik. Sebuah dokumen XML dipecah-pecah berdasarkan elemen yang akan diindeks. Setelah diindeks, dilakukan sejumlah ujicoba untuk setiap jenis indeks.

Keluaran ujicoba tersebut berupa daftar dokumen relevan untuk setiap jenis indeks. Keluaran ini berbeda dengan keluaran sistem yang telah dipaparkan sebelumnya, yang berupa campuran berbagai elemen dari dokumen-dokumen dalam koleksi. Hasil penelitian yang mereka lakukan menunjukkan bahwa perolehan informasi pada elemen yang *article* (*full document* atau elemen terbesar) memberikan hasil yang lebih baik. Namun demikian, unit pengindeksan yang lain tidak bisa diabaikan begitu saja apabila informasi di dalamnya dibutuhkan pengguna.

2.3 Undang-undang Republik Indonesia Berformat XML

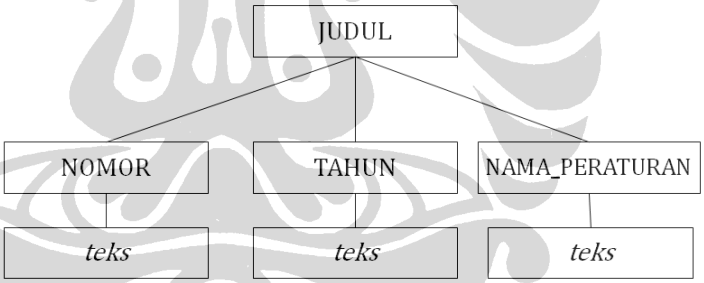
Format XML memberikan kebebasan pada penggunanya untuk merancang sendiri *tag-tag* dan struktur dalam dokumen XML. Pada penelitian ini, format penulisan XML pada Undang-undang Republik Indonesia yang digunakan adalah format yang dirancang oleh Violina. Format penulisan dokumen XML tersebut merujuk pada aturan penulisan perundang-undangan yang diatur oleh Undang-undang Nomor 10 Tahun 2004 Tentang Pembentukan Peraturan Perundang-undangan

[IND2004]. Sementara itu, *tag-tag*, elemen-elemen, dan atribut yang digunakan dalam dokumen undang-undang XML buatan Violina didefinisikan dalam *Document Type Definition* yang disertakan dalam lampiran.



Gambar 2.6. *Root element* dokumen dengan 4 *child element.*

Gambar 2.6 menjelaskan kerangka penulisan undang-undang Republik Indonesia berformat XML. Sebenarnya ada dua elemen lagi yang menjadi pokok dalam kerangka tersebut, yaitu Penjelasan dan Lampiran. Namun, karena sifatnya yang opsional, maka dalam format XML rancangan Violina, dua topik tersebut ditiadakan.

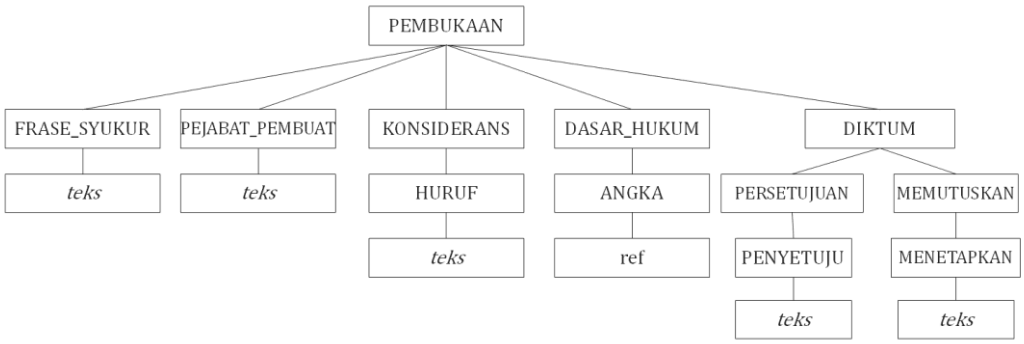


Gambar 2.7. Struktur *tree* pada *node* judul.

Gambar 2.7 menunjukkan struktur *tree* pada node judul. Bagian judul menjelaskan identitas dari undang-undang yang meliputi nomor dan tahun undang-undang serta nama dari ketentuan yang ada diatur oleh undang-undang.

```
<JUDUL>
<NOMOR>11</NOMOR>
<TAHUN>2008</TAHUN>
<NAMA_PERATURAN>
        INFORMASI DAN TRANSAKSI ELEKTRONIK
</NAMA_PERATURAN>
</JUDUL>
```

Gambar 2.8. Contoh elemen judul.



Gambar 2.9. Struktur *tree* dari *node* pembukaan.

Bagian pembukaan seperti Gambar 2.9 terdiri atas frase syukur, jabatan pembentuk peraturan perundang-undangan, konsiderans, dasar hukum pembentukan undang-undang, dan diktum. Frase syukur dalam undang-undang adalah kalimat "DENGAN RAHMAT TUHAN YANG MAHA ESA". Jabatan pembentuk peraturan perundang-undangan adalah frase "PRESIDEN REPUBLIK INDONESIA". Konsiderans adalah poin-poin yang memuat latar belakang dan alasan dibuatnya undang-undang. Konsiderans dimulai dengan kata "Menimbang".

```
<FRASE_SYUKUR>DENGAN RAHMAT TUHAN YANG MAHA ESA</FRASE_SYUKUR>
<PEJABAT_PEMBUAT>PRESIDEN REPUBLIK INDONESIA</PEJABAT_PEMBUAT>
<KONSIDERANS ID="uu-11-2008.konsiderans">
    <HURUF NO="a" ID="uu-11-2008.konsiderans.hr-a">
    bahwa pembangunan nasional adalah suatu proses yang berkelanjutan
    yang harus senantiasa tanggap terhadap berbagai dinamika yang terjadi di
    masyarakat
    </HURUF>
    <HURUF NO="b" ID="uu-11-2008.konsiderans.hr-b">
    bahwa globalisasi informasi telah menempatkan Indonesia sebagai bagian
    dari masyarakat informasi dunia sehingga mengharuskan dibentuknya
    pengaturan mengenai pengelolaan Informasi dan Transaksi Elektronik di
    tingkat nasional sehingga pembangunan Teknologi Informasi dapat
    dilakukan secara optimal, merata, dan menyebar ke seluruh lapisan
    masyarakat guna mencerdaskan kehidupan bangsa
    </HURUF>
</KONSIDERANS>
```

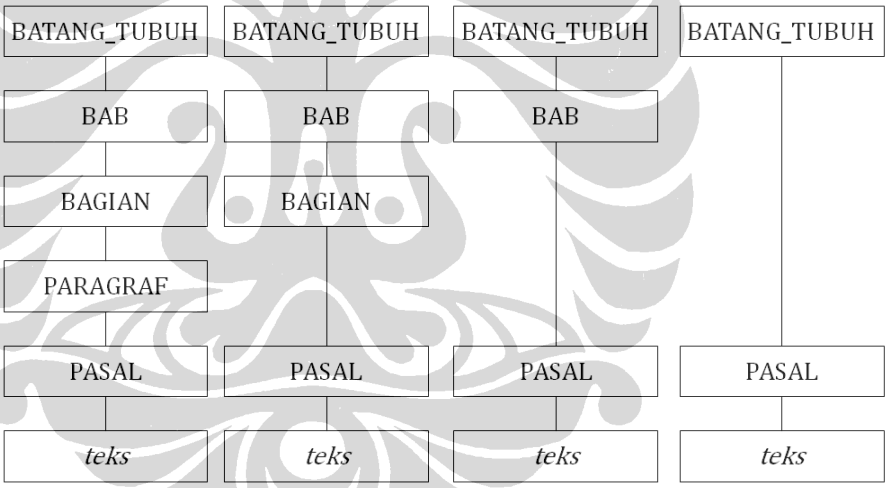
Gambar 2.10. Contoh elemen frase syukur, pejabat pembuat, dan konsiderans.

Dasar Hukum yang dimulai dengan kata "Mengingat" memuat dasar kewenangan pembuatan peraturan perundang-undangan dan peraturan perundang-undangan yang memerintahkan pembuatan undang-undang tersebut. Diktum terdiri atas kata "Memutuskan", "Menetapkan", dan nama dari peraturan perundang-

undangan. Selain itu, diktum ini diawali dengan kalimat "Dengan persetujuan bersama ..." [IND04].

```
<DASAR_HUKUM ID="uu-11-2008.dasar_hukum">
        <REF ID="uud45.ps-5.ay-1"/> dan <REF ID="uud45.ps-20"/>
</DASAR_HUKUM>
<DIKTUM>
<PERSETUJUAN>
        <PENYETUJU NO="1">DEWAN PERWAKILAN RAKYAT REPUBLIK
        INDONESIA</PENYETUJU>
        <PENYETUJU NO="2">PRESIDEN REPUBLIK INDONESIA</PENYETUJU>
</PERSETUJUAN>
<MEMUTUSKAN>
        <MENETAPKAN>
        UNDANG-UNDANG TENTANG INFORMASI DAN TRANSAKSI ELEKTRONIK
        </MENETAPKAN>
</MEMUTUSKAN>
</DIKTUM>
```

Gambar 2.11. Contoh elemen dasar hukum, diktum, dan memutuskan.

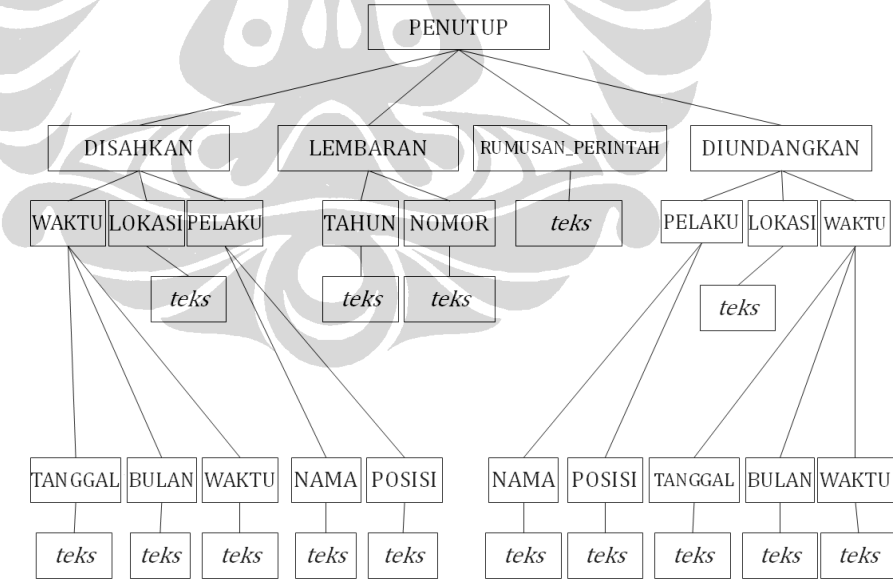


Gambar 2.12. Empat variasi penulisan batang tubuh.

Bagian batang tubuh undang-undang memuat semua substansi peraturan perundang-undangan yang dirumuskan dalam pasal-pasal [IND04]. Gambar 2.12 menunjukan beberapa variasi penulisan bagian batang tubuh undang-undang berformat XML yang ada dalam koleksi. Format yang umum ditemukan pada undang-undang non perubahan adalah batang tubuh dengan bab dan pasal di bawahnya. Sementara itu, pada undang-undang perubahan, bagian yang sering ditemukan adalah batang tubuh dan pasal saja.

```
<BAB NO="IV" ID="uu-11-2008.bb-IV">
<JUDUL_BAB>
PENYELENGGARAAN SERTIFIKASI ELEKTRONIK DAN SISTEM
ELEKTRONIK</JUDUL_BAB>
        <BAGIAN NO="Kesatu" ID="uu-11-2008.bb-IV.bg-Kesatu">
        <JUDUL_BAGIAN>
        Penyelenggaraan Sertifikasi Elektronik</JUDUL_BAGIAN>
        <PASAL NO="13" ID="uu-11-2008.ps-13">
                <AYAT NO="1" ID="uu-11-2008.ps-13.ay-1">
                Setiap Orang berhak menggunakan jasa Penyelenggara
                Sertifikasi Elektronik untuk pembuatan Tanda Tangan
                Elektronik</AYAT>
                <AYAT NO="2" ID="uu-11-2008.ps-13.ay-2">
                Penyelenggara Sertifikasi Elektronik harus memastikan
                keterkaitan suatu Tanda Tangan Elektronik dengan
                pemiliknya</AYAT>
                <AYAT NO="3" ID="uu-11-2008.ps-13.ay-3">
                Penyelenggara Sertifikasi Elektronik terdiri atas:
                <HURUF NO="a" ID="uu-11-2008.ps-13.ay-3.hr-a">
                Penyelenggara Sertifikasi Elektronik Indonesia</HURUF>
                <HURUF NO="b" ID="uu-11-2008.ps-13.ay-3.hr-b">
                Penyelenggara Sertifikasi Elektronik asing</HURUF>
                </AYAT>
        </PASAL>
        </BAGIAN>
</BAB>
```

Gambar 2.13. Contoh elemen pada *node* bab.



Gambar 2.14. Struktur *tree* pada *node* penutup.

```
<PENUTUP>
<RUMUSAN_PERINTAH>
Agar setiap orang mengetahuinya, memerintahkan pengundangan <REF ID="uu-11-
2008"/> dengan penempatannya dalam Lembaran Negara Republik Indonesia
</RUMUSAN_PERINTAH>
<DISAHKAN>
        <LOKASI>Jakarta </LOKASI>
        <WAKTU>
        <TANGGAL>21</TANGGAL><BULAN>April</BULAN><TAHUN>2008</TAHUN>
        </WAKTU>
        <PELAKU>
        <POSISI> PRESIDEN REPUBLIK INDONESIA</POSISI>
        <NAMA> DR. H. SUSILO BAMBANG YUDHOYONO </NAMA>
        </PELAKU>
</DISAHKAN>
<DIUNDANGKAN>
        <LOKASI> Jakarta</LOKASI>
        <WAKTU>
        <TANGGAL>21</TANGGAL><BULAN>April</BULAN><TAHUN>2008</TAHUN>
        </WAKTU>
        <PELAKU>
        <POSISI>MENTERI HUKUM DAN HAK ASASI MANUSIA REPUBLIK
        INDONESIA</POSISI>
        <NAMA>ANDI MATTALATA</NAMA>
        </PELAKU>
</DIUNDANGKAN>
<LEMBARAN>
        <TAHUN>2008</TAHUN> <NOMOR>58</NOMOR>
</LEMBARAN>
</PENUTUP>
</DOKUMEN>
```

Gambar 2.15. Contoh penulisan elemen penutup.

Struktur penulisan bagian penutup dalam format XML ditunjukkan oleh Gambar 2.15. Bagian penutup yang merupakan bagian akhir dari undang-undang memuat empat hal yakni [IND04]:

a) rumusan perintah pengundangan dan penempatan undang-undang dalam Lembaran Negara Republik Indonesia.
b) penandatanganan pengesahan atau penetapan.
c) pengundangan peraturan perundang-undangan.
d) akhir bagian penutup.

2.4 Open Source Search Engine

Saat ini ada beberapa *search engine* yang bersifat *open source* beredar secara luas di internet sebagai alternatif bagi para pengembang aplikasi pencarian. Fungsionalitas yang ditawarkan bisa jadi tidak berbeda dengan *commercial search*

engine, tetapi dengan beberapa keuntungan tambahan. Misalnya dapat diperoleh tanpa mengeluarkan biaya (gratis), tidak perlu melakukan *maintenance* secara aktif, dan memungkinkan pengguna untuk melakukan modifikasi sesuai kebutuhan [MID07]. Banyaknya *open source search engine* tentu memberi keleluasaan bagi pengguna untuk memberikan pilihan. Pengguna tinggal memilih *search engine* mana yang paling sesuai dengan kebutuhannya.

Tabel 2.1. Fitur-fitur beberapa *open source search engine*

Search Engine	Storage (f)	Result Excerpt	Stopwords	Filetype (e)	Stemming	Fuzzy Search	Sort (d)	Ranking	Search Type (c)	Indexer Lang. (b)	License (a)
Datapark	2	Ya	Ya	1,2,3	Ya	Ya	1,2	Ya	2	1	4
ht://Dig	1	Ya	Ya	1,2	Ya	Ya	1	Ya	2	1,2	4
Indri	1	Ya	Ya	1,2,3,4	Ya	Ya	1,2	Ya	1,2,3	2	3
IXE	1	Ya	Ya	1,2,3	-	Ya	1,2	Ya	1,2,3	2	8
Lucene	1	-	Ya	1,2,4	Ya	Ya	1	Ya	1,2,3	3	1
MG4J	1	Ya	Ya	1,2	Ya	-	1	Ya	1,2,3	3	6
mnoGoSearch	2	Ya	Ya	1,2	Ya	Ya	1	Ya	2	1	4
Namazu	1	Ya	-	1,2	-	-	1,2	Ya	1,2,3	1	4
Omega	1	-	Ya	1,2,4,5	Ya	-	1	Ya	1,2,3	2	4
OmniFind	1	Ya	Ya	1,2,3,4,5	Ya	Ya	1	Ya	1,2,3	3	5
OpenFTS	2	-	Ya	1,2	Ya	Ya	1	Ya	1,2	4	4
SWISH-E	1	-	Ya	1,2,3	Ya	Ya	1,2	Ya	1,2,3	1	4
SWISH++	1	-	Ya	1,2	Ya	-	1	Ya	1,2,3	2	4
Terrier	1	-	Ya	1,2,3,4,5	Ya	Ya	1	Ya	1,2,3	3	7
WebGlimpse	1	Ya(g)	-	1,2	-	Ya	1(e)	Ya	1,2,3	1	8,9
XMLSearch	1	-	Ya	3	-	Ya	3	-	1,2,3	2	8
Zettair	1	Ya	Ya	1,2	Ya	-	1	Ya	1,2,3	1	2

(a) 1:Apache,2:BSD,3:CMU,4:GPL,5:IBM,6:LGPL,7:MPL,8:Comm,9:Free
(b) 1:C,2:C++,3:Java,4:Perl;5:PHP,6:Tcl
(c) 1:phrase,2:boolean,3:wild card.
(d) 1:ranking,2:date,3:none
(e) 1:HTML,2:plain text,3:XML,4:PDF,5:PS
(f) 1:file,2:database
(g) commercial version only

[MID07]

Middleton dan Baeza Yates membandingkan performa beberapa *search engine* dalam penelitiannya. Penelitian tersebut bertujuan memberikan referensi kepada khalayak dalam memutuskan pilihan *search engine* yang sesuai dengan masalah pencarian yang dihadapi. Ada 17 *search engine* yang dibandingkan dalam penelitian tersebut. Daftar *search engine* serta fitur yang disediakan dapat dilihat pada Tabel 2.1. Pembandingan meliputi empat hal, yakni waktu yang dihabiskan

untuk mengindeks berbagai koleksi dokumen, ukuran dari indeks, waktu yang dibutuhkan untuk memberikan jawaban, serta kualitas dari jawaban yang diberikan.

Tabel 2.2. Hasil ujicoba

Search Engine	*Indexing Time (ms)*	*Index Size %*	*Searching Time (ms)*	*Answer Quality p@5*
Indri	0:15:45	63	19	**0.2851**
IXE	0:31:10	30	19	0.1429
Swish-E	0:19:45	31	45	-
Terrier	0:40:12	52	50	0.2800
XMLSearch	**0:04:44**	**22**	**12**	-

[MID07]

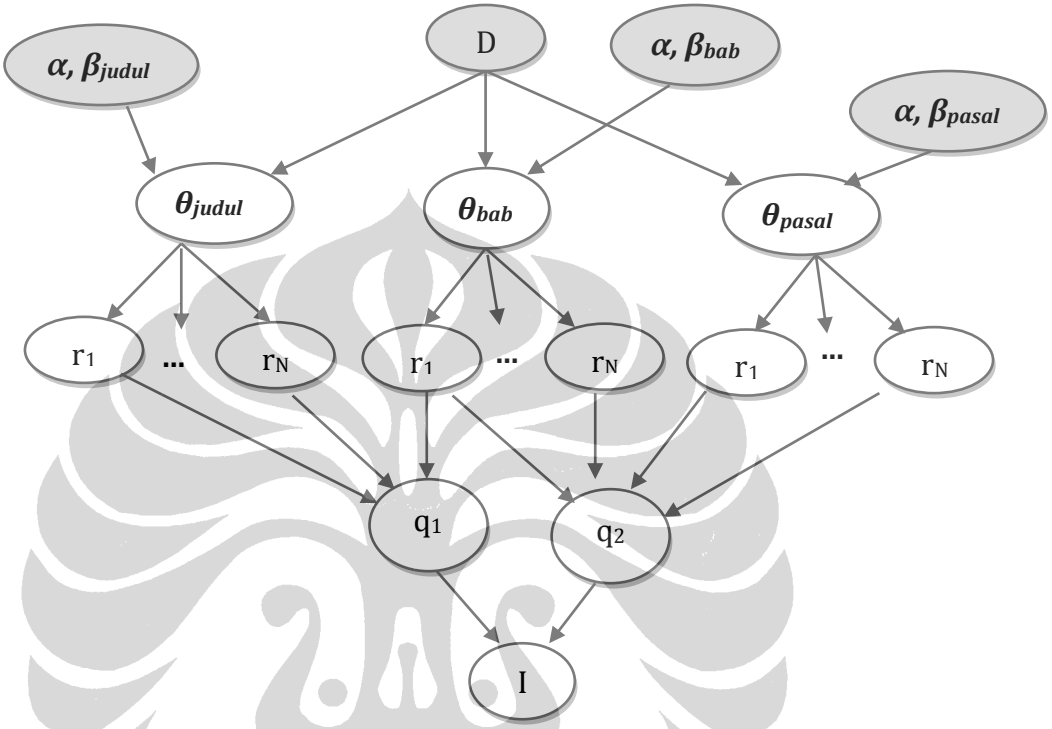
Pada Tabel 2.2, *search engine* XMLSearch menjadi yang terdepan dalam 3 poin perbandingan, yakni waktu pengindeksan, ukuran indeks, dan waktu pencarian. Namun demikian, XMLSearch tidak masuk dalam 3 besar kualitas jawaban yang diukur dengan *precision at 5* karena XMLSearch tidak ada fitur pemeringkatan jawaban. Selain itu, mesin pencari ini tidak memiliki fitur *stemming* dan *ranking*.

Indri *Search Engine* [LEM09], yang merupakan turunan dari Lemur Project, memiliki ukuran indeks yang terbesar dibanding mesin pencari lainnya, tetapi memiliki kemampuan yang di atas rata-rata dalam waktu pengindeksan dan waktu pencarian. Indri hanya kalah oleh XMLSearch yang menduduki peringkat pertama. Selain itu, kualitas jawaban yang diberikannya merupakan yang tertinggi dan fitur yang ditawarkan cukup lengkap. Berdasarkan fakta-fakta tersebut, pada penelitian ini Indri *Search Engine* digunakan sebagai alat bantu dalam proses pengindeksan dokumen dan pencarian informasi.

2.5 Indri Retrieval Model

Model perolehan informasi yang digunakan oleh Indri merupakan gabungan antara *statistical modelling language* dan *inference network*, dua model yang telah secara luas dipelajari dan diterapkan karena terkenal efektif menangani berbagai macam perolehan informasi [MET05]. Kelebihan dua model itulah yang ingin diterapkan pada Indri. Menurut Dai, gabungan dua model yang diterapkan pada

Indri lebih baik dibandingkan model perolehan informasi lain seperti Okapi, KL-divergence, TF*IDF, atau *cosine similarity* [DAI06].



Gambar 2.16. Contoh *inference network* pada Indri.

Gambar 2.16 menunjukkan contoh *inference network* yang digunakan oleh Indri. Gambar tersebut terdiri atas komponen-komponen berikut [MET05]:

a. *Document node* (D), representasi dari dokumen yang diamati.
b. *Model nodes* (θ), yaitu *language model*.
c. *Representasion concept nodes* (r), fitur-fitur pada dokumen.
d. *hyperparameter* (*α, β*), yaitu *smoothing parameter* yang digunakan dalam perhitungan.
e. *belief node* (q), yaitu kueri.
f. *information need node* (I).

2.5.1 *Document node*

Dalam Indri, sebuah dokumen direpresentasikan dalam *multisets of binary vectors*, dimana setiap entri pada vektor biner tersebut merepresentasikan ada atau tidaknya suatu ciri pada teks. Ciri-ciri yang dapat diekspresikan secara biner antara lain apakah suatu kata berada di awal atau di akhir kalimat, berupa *uppercase* atau *lowercase*. Apabila suatu ciri muncul dalam dokumen yang diamati, maka ciri tersebut akan bernilai 1 dan bernilai 0 jika sebaliknya [MET05].

2.5.2 *Model Nodes*

Network pada Gambar 2.16 menunjukkan tiga model yang digunakan untuk merepresentasikan bagian berbeda dari dokumen yang sama. Model judul merepresentasikan potongan dokumen yang terdiri dari semua teks yang terdapat pada bagian judul, sementara model bab terdiri dari semua teks yang terdapat pada bagian bab dokumen sebenarnya.

2.5.3 *Representation of concept nodes*

Representation of concept nodes merupakan ciri-ciri pada dokumen yang diekspresikan sebagai sebuah *binary random variables.* Sebuah ciri pada dokumen bisa muncul berkali-kali dalam *network* dengan induk yang berbeda seperti ditunjukkan oleh Gambar 2.16. Pada gambar tersebut, ciri r1...rN, muncul pada ketiga buah model. Hal tersebut bermanfaat untuk membedakan letak kemunculan suatu kata [MET05].

2.5.4 *Belief nodes*

Belief nodes atau kueri menjelaskan bahwa kepercayaan tentang relevansi suatu dokumen bergantung pada representasi dari dokumen. *Node* ini berada di antara *representation of concept nodes* dengan *information needs nodes*. Apabila suatu ciri (r) muncul pada kueri (q), maka akan ada garis yang menghubungkan kedua *node* tersebut, juga sebaliknya. Hal serupa berlaku pada hubungan antara kueri (q) dengan *node* I. Apabila kueri yang diberikan tidak dapat menunjukkan relevansi dokumen, maka tidak ada garis penghubung di antara dua *node* tersebut.

2.5.5 *Information need node*

Berdasarkan Gambar 2.16, tingkat kepercayaan akan relevan atau tidaknya suatu dokumen terhadap kueri dapat ditemukan di *node* I ini. Pada *node* ini, semua fakta-fakta yang ada dalam *network* dikumpulkan menjadi satu nilai saja [MET05]. Nilai inilah yang akan digunakan dalam proses pemeringkatan hasil pencarian.

2.6 Metode Evaluasi Sistem Perolehan Informasi

Banyak pilihan teknik yang dapat digunakan dalam perolehan informasi. Setiap teknik mempunyai ciri khas atau karakteristik tersendiri, mempunyai tingkat efektifitas yang berbeda untuk kasus yang berbeda. Diperlukan suatu metode evaluasi untuk mengukur tingkat efektifitas teknik perolehan informasi. Evaluasi dilakukan terhadap hasil perolehan informasi. Secara umum, untuk melakukan evaluasi diperlukan tes koleksi yang terdiri dari tiga hal, yakni [MAN08]:

1. Sebuah koleksi dokumen.
2. Sekumpulan topik ujicoba yang merepresentasikan kebutuhan informasi.
3. Satu set *relevance judgements*, yaitu semacam daftar seluruh dokumen dalam koleksi yang relevan terhadap setiap topik ujicoba.

Metode dasar evaluasi sistem perolehan informasi yang umum digunakan yaitu *precision* dan *recall*. Misalkan R adalah kumpulan dokumen yang relevan, q adalah kueri, A adalah dokumen yang dikembalikan oleh sistem saat diberikan kueri q, dan Ra adalah dokumen relevan yang dikembalikan sistem. *Precision* dan *Recall* dapat didefinisikan sebagai berikut:

a) *Recall* adalah perbandingan antara dokumen relevan yang sudah dikembalikan sistem dengan jumlah seluruh dokumen relevan yang ada dalam sistem.
b) *Precision* adalah perbandingan antara dokumen relevan yang sudah dikembalikan sistem dengan jumlah seluruh dokumen yang sudah dikembalikan sistem.

$$Recall = \frac{Ra}{R} \quad Precision = \frac{Ra}{A}$$

Pada metode *precision recall*, dibutuhkan penilaian biner terhadap setiap dokumen yang dikembalikan: relevan atau tidak relevan. Selain itu, dibutuhkan informasi yang lengkap tentang semua dokumen relevan yang ada dalam koleksi. Untuk koleksi dokumen yang kecil, hal itu masih mungkin, tetapi untuk koleksi yang sangat besar sulit untuk mengetahui seberapa banyak jumlah dokumen yang relevan.

Metode pengukuran lain yang merupakan turunan dari *Precision* dan *Recall* antara lain *precision at n* (P@n) dan *mean average precision* (MAP). *Precision at n* adalah rasio jumlah dokumen relevan yang dikembalikan sistem dalam *n* dokumen. Sementara itu, MAP merupakan *mean* dari *average precision* setiap kueri. Artinya apabila ada dalam koleksi ada sepuluh buah kueri, maka setiap kueri dihitung dahulu *average precision*-nya, setelah itu jumlahkan *average precision* 10 kueri tersebut. Terakhir, bagi hasil penjumlahan *average precision* dengan banyaknya kueri, yakni 10. Ilustrasi untuk metode *precision at n* dan MAP ditunjukkan oleh Gambar 2.17.

	Q1	P	Q2	P	Q3	P
1	R	1	R	1	R	1
2	R	1	R	1	R	1
3	T		R	1	R	1
4	R	3/4	R	1	R	1
5	T		T		T	
6	T		T		T	
7	T		R	5/7	T	
8	T		R	6/8	T	
9	T		T		T	
10	R	4/10	T		R	5/10

R = Relevan T = Tidak relevan Q=kueri
P = *precision*

Precision at 5
Q1: 3/5
Q2: 4/5
Q3: 4/5
Rata-rata: 11/15=0.73

Precision at 10
Q1: 4/10
Q2: 6/10
Q3: 5/10
Rata-rata: 15/30=0.5

Average Precision
Q1: (1+1+3/4+4/10)/4 =0.78
Q2: (1+1+1+1+5/7+6/8)/6 =0.87
Q3: (1+1+1+1+5/10)/5 =0.9

MAP
(0.78+0.87+0.9)/3 =0.85

Gambar 2.17. Contoh perhitungan P@5, P@10, dan MAP

Precision at n (p@n) mudah untuk diinterpretasikan [BUC04]. Misalnya apabila nilai p@10 suatu sistem adalah 0,6 berarti dapat satu-satunya interpretasi yaitu 6 dari 10 dokumen teratas merupakan dokumen yang relevan. Selain itu, p@n

merepresentasikan kualitas dari jawaban, karena seringkali pengguna hanya diberikan *n* dokumen pertama dari pencarian, bukan daftar keseluruhan hasil pencarian [MID07]. Kelemahan metode ini yakni bukan pembeda yang *powerful* antar beberapa metode perolehan informasi karena hanya merepresentasikan seberapa banyak dokumen yang berada atau tidak berada dalam *n* dokumen teratas [BUC04]. Pemotongan hasil pada *n* dokumen saja juga menyebabkan p@n memiliki marjin error lebih besar ketimbang MAP.

Sementara itu, MAP merupakan diskriminator yang baik dan menawarkan marjin error yang lebih rendah dibandingkan dengan *precision at n* [MAN08]. Selain itu, metode ini tidak memerlukan informasi tentang semua dokumen relevan dalam koleksi. Kelemahan dari metode ini yaitu hasilnya tidak mudah diinterpretasikan. Nilai MAP 0,6 dapat diinterpretasikan dengan berbagai cara.

2.7 Pembahasan

Tujuan penelitian ini adalah merancang sistem perolehan informasi dokumen legal dengan korpus berupa koleksi undang-undang Republik Indonesia. Undang-undang tersebut merupakan undang-undang yang sudah diberi *tag* XML melalui Sistem Ekstraksi Informasi rancangan Violina. Elemen dokumen yang diambil sebagai unit pengindeksan adalah keseluruhan dokumen, elemen bab, dan elemen pasal.

Fitur yang digunakan saat pembuatan indeks yaitu menggunakan proses penghilangan *stopwords* dan tanpa proses *stemming*. Proses pengindeksan dan perolehan informasi menggunakan Indri versi 2.8. Model perolehan informasi yang digunakan *search engine* tersebut adalah *Indri model* yang merupakan gabungan dari *statistical language modelling* dan *inference network*. Kueri yang digunakan adalah kueri teks, bertipe boolean, tanpa proses *stemming*, dan tanpa penghilangan *stopwords*. Penelitian ini menggunakan metode evaluasi MAP, P@5, dan P@10. MAP dipilih karena kelebihannya yang mampu membedakan kualitas antar metode, P@N dipilih karena mempresentasikan kualitas jawaban, yakni seberapa banyak dokumen relevan dalam *n* dokumen teratas.