BASISDATA TEKSTUAL: STUDI KASUS PENGEMBANGAN BASISDATA KAMUS BESAR BAHASA INDONESIA

Wahyu C. Wibowo
Fakultas Ilmu Komputer - Universitas Indonesia

I. Penyimpanan data KBBI

Data Kamus Besar Bahasa Indonesia (KBBI) memiliki karakteristik yang khusus, yakni panjang setiap entri yang sangat bervariasi antara satu entri dengan entri yang lain. Hal ini menimbulkan masalah bila data KBBI akan disimpan dan diproses menggunakan fasilitas sistem Pengelola Basisdata Relasional (*Relational Database Management System*/RDBMS) seperti xBase. Sistem Pengelola Basisdata tersebut hanya cocok dipergunakan untuk menangani data *Record* dengan panjang yang tetap seperti data Kemahasiswaan atau Kepegawaian.
Menyimpan data dengan panjang bervariasi pada RDBMS dapat dilakukan dengan:

1. Menyimpan dalam *field* dengan panjang tetap dengan kemungkinan menampung data yang paling panjang
2. Menyimpan dalam sistem penyimpan data khusus yang tersedia seperti *Variable Length Character field*, *Memo Field*, atau *Binary Large Object Field*.

Cara pertama mengharuskan kita untuk memperkirakan panjang data maksimum yang akan disimpan dan menyediakan tempat penyimpan data dengan panjang yang maksimum ini. Apabila variasi panjang data sangat beragam seperti pada KBBI, maka akan banyak tempat penyimpan yang terbuang tidak terpakai. Selain itu, panjang maksimum yang diperkirakan pada suatu saat mungkin tidak lagi *valid* pada saat yang lain sehingga diperlukan reorganisasi seluruh data untuk menangani data yang lebih panjang lagi.

Cara kedua memungkinkan penyimpanan data tanpa perlu mengetahui ukuran data terpendek atau terpanjang, semua akan dapat ditampung oleh RDBMS. Meskipun demikian, pada umumnya efektivitas simpan

atau proses tidak dapat diperoleh karena penggunaan tempat simpan dengan ukuran kelipatan tertentu (misal kelipatan ukuran sektor) atau penyimpanan dengan strategi penyimpanan khusus yang memerlukan waktu akses yang signifikan. Sebagai contoh, penggunaan field Memo pada xBase memerlukan tempat berukuran kelipatan 512 karakter. Apabila rata-rata separuh dari ukuran tempat 512 karakter tidak terisi dan dimiliki 100.000 data, maka akan terbuang tempat simpan berukuran 25.600.000 karakter. Data untuk field sejenis ini umumnya juga disimpan terpisah dengan data field umum lainnya (karakter, numerik, *date*, dll) sehingga memerlukan waktu ekstra untuk mengakses data.

Saat ini, Perangkat Lunak khusus pengelola data tekstual dapat ditemukan di pasar dengan berbagai fasilitas dan kelebihannya. Satu hal yang membuat alternatif ini tidak begitu menarik: Harganya yang tidak murah.

II. Pemakaian KBBI

Dari segi pemakaian, pemakai KBBI dapat dikelompokkan pada dua golongan pemakai, yaitu:

1. Pembangun dan Perawat KBBI
2. Pengguna KBBI

Pengguna KBBI hanya memerlukan fasilitas untuk membaca atau mencari data, tanpa dapat melakukan perubahan isi KBBI. Pembangun dan perawat KBBI di lain pihak, memerlukan berbagai fasilitas untuk dapat menambahkan entri baru, mengubah entri isi, atau menghapus entri yang sudah ada selain kemampuan dasar untuk membaca atau mencari entri yang sudah ada.

Dari jenis penggunaan tersebut, jelaslah bahwa penyediaan fasilitas akses KBBI kepada publik dapat dilakukan dengan kemudahan tertentu tanpa perlu menimbang kemungkinan masalah perubahan data. Kemampuan fasilitas khusus hanya diperlukan oleh para Pembangun dan perawat KBBI.

III. Pengembangan Sistem Penyimpan Data Khusus untuk KBBI

Salah satu alternatif untuk menyimpan data KBBI adalah dengan membangun sendiri basisdata tekstual untuk KBBI. Sistem ini haruslah mengatasi masalah pemborosan tempat simpan, mengatasi masalah kecepatan akses, serta mudah dioperasikan dan dikembangkan. Dengan cara ini, sistem KBBI dapat diimplementasikan ke berbagai jenis perangkat keras dengan berbagai sistem operasi untuk menunjang pemasyarakatan KBBI.

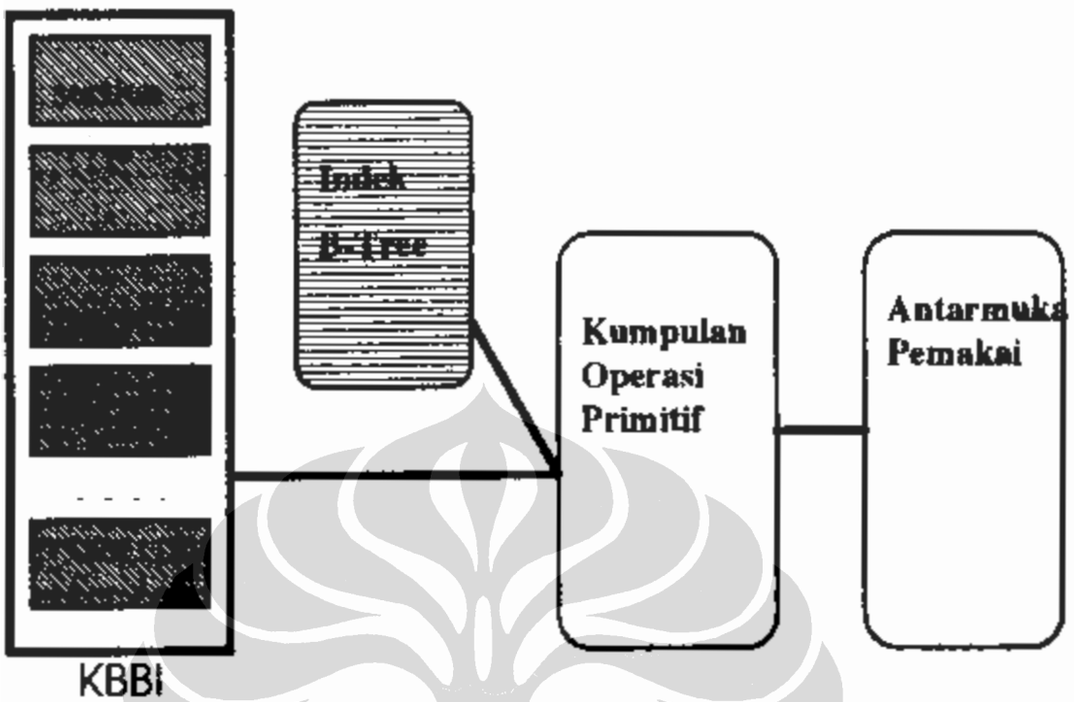
Efisiensi tempat simpan dapat dicapai dengan menggunakan tempat simpan sebagaimana yang diperlukan dengan *overhead* yang minimum. Karena isi KBBI umumnya berupa data teks, efisiensi tempat simpan dapat ditingkatkan lebih lanjut dengan melakukan kompresi isi data.

Kemampuan akses entri yang cepat dapat diperoleh dengan penggunaan teknik peng-indek-an data. Metoda menggunakan keluarga B-Tree adalah metoda yang paling umum untuk data besar.

Agar sistem mudah dikembangkan dan disesuaikan ke berbagai kebutuhan, dibuat sistem dua lapis untuk melakukan operasi pada data beserta indek nya. Lapisan pertama adalah primitif-primitif operasi yang tersedia untuk berinteraksi langsung dengan data dan indek. Lapisan ke dua adalah lapisan Antarmuka pemakai yang menyediakan cara yang mudah untuk berinteraksi dengan KBBI.

IV. Rancangan Awal Pengembangan

Ide dasar dari sistem basisdata KBBI adalah sekumpulan **bagian** (*section*) yang menyimpan data satu **kosa kata** serta file indek yang digunakan untuk mengakses data.



Setiap bagian akan diidentifikasikan oleh posisi (alamat/*offset*) bagian tersebut serta simbol pemberi tanda akhir dari satu bagian/*section* di akhir bagian tersebut. Identifikasi alamat akan disimpan dalam file indek B-Tree. Isi teks suatu bagian yang memerlukan perlakuan khusus (kosa kata, sinonim, antonim, dll) akan diidentifikasi secara khusus agar mudah direferensi oleh program aplikasi. Contoh bentuk identifikasi ini misalnya:
Kosa kata:

αkosa_kataβ atau
γkosa_kataβ untuk identifikasi bagian yang sudah tidak aktif lagi (telah diubah/hapus)

Sinonim, Antonim, dll:

ωkataψ

- Akses/Pencarian/*search*
  Dilakukan dengan parameter nama kosa kata yang akan dicari kemudian dilakukan pencarian posisi data pada B-Tree dan dilakukan akses data bila ditemukan.

- Penambahan Kosa Kata Baru
  Dilakukan dengan menambahkan bagian baru ke akhir dokumen.

- Pengubahan/Penghapusan data Kosa Kata
  Dilakukan dengan menandai bagian sebagai "tidak aktif" dan menambahkan bagian baru yang telah mengalami perubahan di akhir dokumen.

- Reorganisasi
  Diperlukan untuk membuang semua bagian yang tidak aktif dan membangun kembali struktur B-Tree.

V. Fasilitas Lanjut

Agar sistem ini memiliki fungsionalitas yang baik sebagai Basisdata tekstual, perlu ditambahkan berbagai fasilitas penunjuang seperti:

Pengendali Transaksi
Diperlukan untuk menjamin kebenaran data saat dilakukan operasi terhadap data. Operasi terhadap data dapat gagal apabila terjadi misalnya karena terjadi pelanggaran terhadap aturan integritas data, pemakai membatalkan operasi, atau sistem mengalami crash.

Pengendali Akses Simultan (*concurrency control*)
Bila sistem KBBI akan dirawat oleh sejumlah orang pada suatu saat, diperlukan mekanisme pengaturan akses agar diperoleh kebenaran hasil operasi.

Backup dan Recovery
Fasilitas ini disediakan untuk melakukan *backup* data KBBI beserta informasi indek yang terkait dan untuk melakukan *restore* data backup bila terjadi kegagalan sistem.

Pengendali Akses Data
Sistem pengamanan diperlukan untuk memberi otorisasi akses kepada sejumlah pemakai. Otorisasi diberikan dalam bentuk Hak Baca, Hak Ubah, Hak Tambah, atau Hak Hapus data KBBI.

VI. Kesimpulan

- Pemasyarakatan KBBI dapat ditingkatkan dengan membangun sendiri pengelola basisdata KBBI.

- Pengalaman dalam membangun pengelola basisdata KBBI dapat diterapkan untuk membangun basisdata teks lain seperti basisdata Berita, Peraturan Perundangan, atau Naskah-naskah tekstual lainnya.