**Intelijen dalam Cloud**

**7.1 Pendahuluan**

Kami telah melihat bagaimana teknologi web dan cloud memungkinkan kami untuk menyimpan dan memproses dengan mudah jumlah informasi, dan seperti yang kita lihat di Chap. 6, ada banyak penyimpanan data yang berbeda model yang digunakan di cloud. Bab ini melihat bagaimana kita bisa mulai membuka kunci yang tersembunyi potensi data itu, untuk menemukan ‘nugget emas’ dari informasi yang benar-benar berguna dalam kumpulan sampah yang tidak relevan atau tidak berguna dan untuk menemukan pengetahuan baru melalui analisis cerdas data. Banyak alat dan teknik cerdas dis- Disumpahi di sini berasal baik sebelum cloud computing. Namun, sifat dari data cloud, aksesnya yang terukur ke sumber daya besar dan ukuran data yang tersedia berarti itu kelebihan alat-alat ini jauh lebih jelas. Kami sekarang telah mencapai suatu tempat di mana banyak tugas berbasis web yang umum tidak akan mungkin tanpa mereka. Sebagian besar informasi baru ini datang langsung dari pengguna. Seni penyadapan ke dalam kedua data yang dibuat oleh pengguna dan interaksi pengguna dengan web dan cloud- aplikasi berbasis membawa kita ke medan kecerdasan kolektif dan ‘crowourcing sourcing’.

Kecerdasan kolektif memiliki akar yang mendahului web di berbagai bidang termasuk biologi dan sosiologi. Penerapan teknik kecerdasan kolektif untuk data berbasis web telah menerima banyak perhatian dalam beberapa tahun terakhir, dan itu telah menjadi jelas bahwa, khususnya dengan ekspansi sistem cloud yang cepat, ini adalah daerah yang subur untuk mengembangkan cara baru mengakses, mensintesis dan, secara cerdas menganalisis data kami. Telah disarankan bahwa hasil akhirnya adalah Web 3.0.

Kami akan memulai bab ini dengan gambaran singkat tentang jenis data yang kami hadapi dengan dan teknik yang sudah digunakan untuk mengekstrak intelijen dari data pengguna, interaksi, dan kolaborasi. Kami kemudian akan melihat beberapa detail di proses pencarian termasuk penilaian tantangan menghadapi tekstual data dan ikhtisar prosedur yang mendasari mesin pencari. Kami kemudian pindah ke bidang kecerdasan kolektif dan apa yang dapat ditawarkan aplikasi web dan cloud tions. Kami akan menyelesaikannya dengan melihat kekuatan visualisasi. Di akhir ini bab, latihan akan menggunakan pustaka sumber terbuka untuk melakukan beberapa tahap mengekstraksi dan menganalisis teks online.

**7.2 Web 2.0**

Sepuluh tahun yang lalu sebagian besar situs web memiliki nuansa kuliah: Informasi statis mengalir satu arah dari situs ke pengguna. Sebagian besar situs sekarang dinamis dan interaktif, dan hasilnya lebih baik digambarkan sebagai percakapan. Kita mungkin harus melakukannya ingat bahwa HTTP itu sendiri dirancang untuk bekerja dengan cara percakapan. Web 2.0 adalah semua tentang mengizinkan dan mendorong pengguna untuk berinteraksi dengan situs web. Sebagai interaksi peningkatan tion, demikian juga jumlah data. Dalam bab ini, kita akan melihat beberapa alat yang dapat kami terapkan untuk mendapatkan lebih banyak konten buatan pengguna. Seringkali pro Penggunaan konten ini secara efektif untuk meningkatkan situs kami mengarah pada peningkatan minat dan aktivitas lebih lanjut sehingga menciptakan lingkaran berbudi luhur.

**7.3 Relational Databases**

Teknologi database relasional sudah matang dan dikenal oleh sebagian besar pengembang perangkat lunak dan manfaat dari bahasa SQL yang ringkas dan kuat. Data disimpan dalam database relasional memiliki sejumlah keunggulan yang berbeda ketika datang ke pengambilan informasi termasuk:

• Data disimpan dalam bidang berlabel.

• Field el (atau kolom) memiliki tipe data dan atribut ukuran yang telah ditentukan sebelumnya.

• Kita dapat menetapkan batasan pada data akhir, misalnya, 'semua nilai yang dimasukkan harus unik ',' nilai nol tidak diterima 'atau' data yang dimasukkan harus berada di dalam satu set atau rentang nilai tertentu ’.

• Teknik normalisasi yang dipahami dengan baik dapat diterapkan yang telah dilakukan

terbukti mengurangi redundansi dan memberikan pemahaman logis yang mudah kepada fisi pemetaan penyimpanan cal.

**7.4 Data Teks**

Meskipun manfaat yang tercantum di atas, telah diperkirakan bahwa sekitar 80% dari data organisasi dalam format tidak terstruktur. Datanya tentu saja multimedia, tetapi data teks telah menarik minat yang besar sebagai sumber utama untuk penambangan web meskipun dalam beberapa tahun terakhir perhatian beralih ke format lain, khususnya gambar dan video di mana kemajuan signifikan telah terjadi.

Data teks tidak terstruktur memiliki properti yang berguna sebagai manusia dan mesin dapat dibaca, meskipun tidak masuk akal ke mesin. Aturan untuk data teks sangat berbeda dengan database relasional. Data teks mungkin atau mungkin tidak divalidasi dan sering diduplikasi berkali-kali. Itu mungkin memiliki beberapa struktur, seperti dalam akademik artikel, atau sedikit atau tidak ada struktur, misalnya, dalam hal blog dan email. Di beberapa kasus, ejaan dan tata bahasa diperiksa dengan sangat hati-hati, tetapi dalam kasus lain banyak kesalahan mengambil, salah ejaan, kata-kata slang, singkatan dan akronim adalah hal biasa. Ada banyak yang terkenal untuk banyak hubungan antara kata-kata dan makna. Dalam data teks kami sering menemukan sinonim yang kata-kata yang berbeda dengan arti dan hom- yang sama

onyms yang kata-kata dengan ejaan yang sama tetapi dengan arti yang berbeda. Contoh

ditunjukkan pada Tabel 7.1 dan 7.2.

Teks juga memiliki banyak contoh unit multi-kata seperti 'pengambilan informasi' di mana lebih dari satu kata mengacu pada konsep tunggal, kata-kata baru terus-menerus Muncul, ada banyak contoh crossover multinasional dan arti dari kata-kata dapat bervariasi dari waktu ke waktu. Untuk alasan ini mengekstraksi informasi yang berguna dari data tekstual dalam banyak hal merupakan masalah yang lebih sulit daripada dengan data yang disimpan dalam relasional database. Contoh-contoh khas dari teks yang mungkin ditargetkan untuk analisis cerdas adalah artikel, kertas putih, informasi produk, ulasan, blog, wiki dan pesan papan. Salah satu efek dari Web 2.0 telah sangat meningkatkan jumlah teks online data. Cloud menyediakan akses mudah ke kekuatan komputasi yang dapat diukur memproses data ini dengan cara yang inovatif dan bermanfaat. Web itu sendiri dapat dianggap sebagai satu toko data besar. Kami akan melihat alat dan teknik yang telah dikembangkan oped untuk memaksimalkan potensi sumber daya yang fantastis ini di ruang terbuka manusia.

**7.5 Pengolahan Bahasa Alami**

Pemrosesan bahasa alami adalah serangkaian teknik yang digunakan untuk memproses tertulis dan bahasa manusia lisan. Tugas pemrosesan bahasa alami sering melibatkan pengkategorisan jenis kata yang terjadi dalam teks. Jenis data yang sangat berguna mengacu pada hal-hal seperti negara, organisasi, dan individu. Proses identifikasi otomatis ini dikenal sebagai ekstraksi entitas. Tugas umum kedua adalah mengidentifikasi 'bagian dari ucapan' dari kata-kata tertentu sehingga, misalnya, kata benda, kata sifat dan keterangan dapat secara otomatis diidentifikasi. GATE (http://gate.ac.uk/) tersedia secara gratis alat sumber terbuka yang bisa ditulis di Java yang melakukan tugas di atas bersama lebih banyak terkait dengan pemrosesan teks. Pemrosesan bahasa alami telah berhasil kemajuan di bidang-bidang seperti terjemahan otomatis, pengenalan suara dan tata Bahasa memeriksa. Salah satu tujuan jangka panjang dari pemrosesan bahasa alami adalah untuk memungkinkan mesin untuk benar-benar memahami teks manusia, tetapi masih ada tantangan besar diatasi, dan kita harus berhati-hati dalam kasus klaim yang dimiliki oleh tujuan ini telah dicapai oleh sistem yang sudah ada atau bahwa solusinya sangat dekat.

**7.6 Pencarian**

Pencarian adalah salah satu aplikasi pertama yang kita mungkin lampirkan istilah 'Cloud', dan pencarian tetap, bagi sebagian besar pengguna, alat paling penting di web. Mesin pencari seperti Google, Bing, dan Yahoo adalah contoh terkenal yang matang aplikasi cloud dan juga memberikan pengenalan yang baik untuk banyak konsep penting berkaitan dengan data cloud bersama dengan rekayasa teks dan kecerdasan kolektif. Oleh karena itu, patut dilihat singkat 'di bawah tenda' untuk mendapatkan pengantar inner workings search engine. Serta memberikan wawasan ke salah satu kunci alat untuk data web dan cloud, penyelidikan kami akan mengungkap sejumlah tema dan konsep penting yang sangat relevan dengan web dan cloud intelligence dan memanfaatkan data yang berisik dan tidak terstruktur yang umumnya ditemukan di web. Tentu saja, pasar mesin pencari komersial adalah bisnis besar, dan vendor tetap tepat

kerja sistem mereka tersembunyi dengan baik. Meskipun demikian, ada sejumlah bebas tersedia pustaka sumber terbuka, dan di akhir bab ini, kami membangunnya mesin pencari sendiri. Kami akan menggunakan Apache yang banyak digunakan dan banyak dipuji Indeks Lucene bersama dengan proyek-proyek Apache terkait Nutch, Tika dan Solr. Ini berharga mencatat bahwa Nutch secara khusus dirancang untuk dijalankan pada implementasi Apache Hadoop MapReduce yang kami selidiki di Chap.

**7.6.1 Ikhtisar Mesin Telusur**

Mesin pencari memiliki tiga komponen utama seperti yang ditunjukkan pada Gambar 7.1:

1. Crawler

2. Pengindeks

3. Antarmuka web

**7.6.2 Perayap**

Web tidak seperti database lain yang pernah ditemukan sebelumnya dan memiliki banyak keunikan fitur yang membutuhkan alat baru serta adaptasi alat yang ada. Sedangkan Lingkungan sebelumnya seperti database relasional menyarankan bahwa pencarian harus mencakup seluruh database dengan cara yang tepat dan dapat diprediksi, ini sama sekali tidak mungkin ketika kami mengukur skala web. Perayap web (sering disebut spider atau bot) menjelajah web dengan cara metodis sering dengan tujuan untuk menutupi fraksi signifikan dari bagian yang tersedia untuk publik dari seluruh web. Untuk memulai merangkak, satu set halaman benih yang baik diperlukan. Cara umum untuk melakukan ini adalah mendapatkan benih dari proyek seperti Open Directory Project (http: // www.dmoz.org/). Setelah daftar halaman benih telah diperoleh, pada dasarnya crawler bekerja dengan melakukan serangkaian perintah HTTP GET (lihat Bab 4). Hyperlink ditemukan di halaman apa pun yang disimpan dan ditambahkan ke daftar situs yang akan diambil. Tentu saja, ada banyak komplikasi seperti penjadwalan merangkak, parallelisation, prioritisation, kesopanan (menghindari situs luar biasa tertentu dengan menghormati mereka kebijakan terkait perayapan) dan penanganan duplikat dan tautan mati. Ada sejumlah pustaka sumber terbuka yang dapat digunakan untuk perayapan mulai dari cukup sederhana dan mudah digunakan HTTP Track (http://www.httrack.com/) ke Apache Nutch (http://nutch.apache.org/) yang dapat digunakan untuk membangun keseluruhan pencarian web mesin. Kami melihat ini di akhir tutorial bab kami. Merangkak seluruh web adalah usaha besar yang membutuhkan sumber daya besar dalam hal penyimpanan, komputasi dan manajemen. Dalam kebanyakan kasus, organisasi akan melakukan perayapan mereka memiliki data sendiri atau mungkin perayapan web yang terfokus atau cerdas pada area topik tertentu.

**7.6.3 The Indexer**

Alih-alih mencoba menjawab pertanyaan 'apa kata-kata yang terkandung dalam tertentu dokumen? ’, yang dapat Anda jawab hanya dengan membaca dokumen, pengindeks bertujuan untuk memberikan jawaban cepat atas pertanyaan 'dokumen apa yang berisi kata khusus ini?', dan untuk alasan ini indeks disebut sebagai 'indeks terbalik'. Setelah perayap mengembalikan laman, kami perlu memprosesnya dengan cara tertentu. Halaman bias berada dalam berbagai format seperti HTML, XML, Microsoft Word, Adobe PDF dan teks biasa. Umumnya, tugas pertama dari pengindeks adalah mengekstrak data teks

dari berbagai format yang mungkin ditemui. Ini adalah tugas yang rumit tapi untungnya perangkat open source seperti Apache Tika (http://tika.apache.org/) bebas tersedia. Setelah teks diekstraksi, kita dapat mulai membangun indeks dengan memproses aliran teks. Sejumlah proses biasanya dilakukan sebelum kami membangun Indeks. Perhatikan bahwa langkah-langkah ini akan mengurangi jumlah kata yang disimpan mungkin membantu ketika mengekstraksi kecerdasan dari data teks

**7.6.3.1 Tokenisasi**

Tugas dari tokeniser adalah memecah aliran karakter menjadi kata-kata. Dalam Bahasa Inggris atau sebagian besar bahasa Eropa, ini cukup sederhana sebagai karakter ruang dapat digunakan untuk memisahkan kata-kata. Seringkali tanda baca dihapus selama tahap ini. Di bahasa seperti bahasa Tionghoa di mana tidak ada padanan langsung dari karakter luar angkasa, ini adalah tugas yang jauh lebih menantang.

**7.6.3.2 Diatur ke Huruf Kecil**

Ketika program komputer membaca data teks, bentuk huruf besar dan kecil dari indi karakter vidual diberi kode terpisah, dan oleh karena itu, dua kata seperti ‘cloud’ dan ‘Cloud’ akan diidentifikasi sebagai kata-kata terpisah. Secara umum berguna untuk mengatur semuanya karakter ke bentuk standar sehingga kata yang sama dihitung apakah, untuk Misalnya, di awal atau tengah kalimat meskipun beberapa informasi semantik mungkin hilang.

**7.6.3.3 Menghentikan Penghapusan Word**

Seringkali berguna untuk menghapus kata-kata yang sangat sering dalam teks tetapi yang membawa

nilai semantik rendah. Misalnya, Apache Lucene (<http://lucene.apache.org>) sistem pengindeksan berisi daftar kata-kata berhenti berikut ini:

‘A’, ‘an’, ‘dan’, ‘are’, ‘as’, ‘at’, ‘be’, ‘but’, ‘oleh’, ‘Untuk’, ‘if’, ‘in’, ‘into’, ‘is’, ‘it’, ‘tidak’, ‘not’, ‘of’, ‘On’, ‘or’, ‘such’, ‘that’, ‘the’, ‘their’, ‘then’, ‘there’, 'Ini', 'mereka', 'ini', 'untuk', 'adalah', 'akan', 'dengan' Termasuk, bahkan kumpulan kata-kata kecil ini dapat sangat mengurangi jumlah total kata-kata yang disimpan dalam indeks.

**7.6.3.4 Stemming**

Stemming memungkinkan kita untuk menyajikan berbagai bentuk kata menggunakan satu kata. Sebagai contoh, jika teks berisi kata ‘proses’, ‘pemrosesan’, dan ‘proses’, kami akan mengubah sider ini sebagai kata yang sama untuk keperluan pengindeksan. Sekali lagi, ini juga bisa sangat kurangi jumlah kata yang disimpan. Berbagai metode untuk stemming telah diusulkan; yang paling banyak digunakan adalah algoritma stemming Porter (Lucene mengandung kelas-kelas untuk melakukan Porter stemming). Namun, ada beberapa kerugian stemming, dan beberapa search engine tidak menggunakan stemming sebagai bagian dari indexing. Jika stemming digunakan saat membuat indeks, proses yang sama harus diterapkan pada kata-kata yang diketik pengguna ke dalam antarmuka pencarian.

**7.6.4 Pengindeksan**

Begitu kata-kata itu telah diekstraksi, disinggung, disaring dan mungkin berasal, mereka dapat ditambahkan ke indeks. Indeks akan digunakan untuk menemukan dokumen yang relevan ke kueri penelusuran pengguna saat dimasukkan ke antarmuka penelusuran. Sebuah indeks biasanya akan dapat dengan cepat mengembalikan daftar dokumen yang mengandung tertentu kata bersama dengan informasi lain seperti frekuensi atau pentingnya itu kata dalam dokumen tertentu. Banyak mesin pencari memungkinkan pengguna untuk secara sederhana masukkan satu atau lebih kata kunci. Atau mereka dapat membangun kueri yang lebih rumit, misalnya, mengharuskan dua kata harus muncul sebelum dokumen dikembalikan atau bahwa kata tertentu tidak terjadi. Lucene memiliki berbagai jenis kueri

tersedia.

**7.6.5 Peringkat**

Mesin pencari yang mengindeks koleksi ratusan dokumen milik organisasi mungkin menghasilkan hasil yang dapat diterima untuk sebagian besar pencarian, terutama jika pengguna mendapatkan keahlian dalam jenis permintaan yang lebih canggih. Namun, bahkan dalam hal ini beberapa kueri mungkin mengembalikan terlalu banyak dokumen agar bermanfaat. Dalam kasus web, itu jumlahnya menjadi luar biasa, dan situasinya hanya diperparah oleh fakta bahwa web tidak memiliki otoritas pusat untuk menerima, mengkategorikan, dan mengelola dokumen;

siapa pun dapat mempublikasikan di web dan data tidak selalu dapat dipercaya. Salah satu yang paling banyak solusi mudah dan banyak digunakan adalah menempatkan hasil kueri secara berurutan, sehingga halaman atau dokumen yang paling mungkin memenuhi persyaratan pengguna berada bagian atas daftar. Ada beberapa cara untuk melakukan ini.

**7.7 Model Ruang Vektor**

Model ruang vektor (VSM) awalnya dikembangkan oleh Salton pada 1970-an adalah a cara yang ampuh untuk menempatkan dokumen dalam rangka relevansi dengan permintaan tertentu. Setiap kata atau istilah yang tersisa setelah penghapusan kata berhenti dan stemming dianggap sebagai 'dimensi', dan masing-masing dimensi diberi bobot. Model tidak memperhitungkan urutan kata dalam dokumen dan kadang-kadang disebut model 'tas kata-kata'. Nilai bobot (W) terkait dengan frekuensi setiap istilah dalam dokumen (yang frekuensi istilah) dan disimpan dalam bentuk vektor. Vektor adalah cara yang bermanfaat

merekam besarnya dan arahnya. Bobot setiap istilah harus diwakilkan mengirim kepentingan relatif dari istilah dalam dokumen. Kedua dokumen (d) dan queries (q) dapat disimpan dengan cara ini:

dj w jw j wt j = ... (1,, 2,,,,)

qw qw qwt q = ... (1,, 2,,,,)

Kedua vektor kemudian dapat dibandingkan dalam ruang multidimensi yang memungkinkan untuk daftar peringkat dokumen yang akan dikembalikan berdasarkan kedekatannya dengan kueri tertentu. Kita bisa hanya menyimpan nilai biner yang menunjukkan ada atau tidaknya sebuah kata dalam dokumen atau mungkin frekuensi kata sebagai nilai berat dalam vektor.

Namun, cara yang populer dan umumnya lebih efektif dalam menghitung nilai-nilai yang disimpan untuk setiap istilah dikenal sebagai pembobotan tf-idf (frekuensi istilah-inverse dokumen frekuensi) yang didasarkan pada dua pengamatan empiris mengenai koleksi teks:

1. Semakin banyak kata terjadi dalam dokumen, semakin relevan untuk topik tersebut dari dokumen.

2. Semakin banyak kata terjadi di seluruh dokumen dalam koleksi, yang lebih buruk itu membedakan antar dokumen.

Ini berguna untuk menggabungkan frekuensi jangka (tf) dan kebalikan dari jumlah dokumen (idf) dalam koleksi di mana istilah tersebut terjadi setidaknya sekali untuk membuat berat.

tf-idf weighting memberikan bobot pada suatu kata dalam dokumen yang proporsional dengan jumlah kemunculan kata dalam dokumen dan dalam proporsi terbalik jumlah dokumen dalam koleksi yang paling sedikit kata itu sekali, yaitu,

w i j tf i j N df i () () ()

Bobot dari istilah i dalam dokumen j adalah frekuensi dari istilah i dalam dokumen j dikalikan dengan log dari jumlah dokumen (N). Log digunakan sebagai cara ‘squashing’ atau mengurangi perbedaan dan dapat dihilangkan tetapi telah terjadi ditemukan untuk meningkatkan efektivitas. Mungkin lebih mudah diikuti dengan sebuah contoh. Asumsikan kita memiliki 500 dokumen dalam koleksi kami dan kata 'cloud' muncul di 60 ini. Pertimbangkan dokumen tertentu di mana kata 'cloud' muncul 11 kali.

Untuk menghitung tf-idf untuk dokumen itu:

Frekuensi istilah (tf) untuk ‘cloud’ adalah 11.

Frekuensi dokumen invers (idf) adalah log (500/60) = 0,92.

Skor tf - idf yang merupakan produk dari kuantitas ini adalah 11 × 0,92 = 10,12.

Kami kemudian dapat mengulangi perhitungan ini di semua dokumen dan masukkan tf – idf nilai-nilai ke dalam vektor istilah. Ada banyak variasi dari rumus di atas yang memperhitungkan yang lain faktor-faktor seperti panjang setiap dokumen. Banyak sistem pengindeksan, seperti Apache Lucene, akan menyimpan vektor istilah menggunakan beberapa bentuk tf - idf pembobotan. Dokumen (atau halaman web) kemudian dapat dikembalikan dengan cepat dalam urutan tergantung pada perbandingan istilah vektor permintaan dan dokumen dari indeks. Terbalik indeks, seperti yang dibuat oleh Apache Lucene, pada dasarnya adalah data yang ringkas struktur di mana representasi vektor jangka dokumen disimpan.

**7.8 Klasiﬁkasi**

Kami menemukan klasi ﬁ kasi di semua bidang usaha manusia. Klasi ﬁ kasi digunakan sebagai

sarana pengorganisasian dan penataan dan umumnya membuat data dari semua jenis dapat diakses dan dikelola. Tidak mengherankan, klasiﬁkasi telah menjadi subyek penelitian intensif dalam ilmu komputer selama beberapa dekade dan telah muncul sebagai sebuah komponen penting dari sistem cerdas.

Sehubungan dengan data teks tidak terstruktur, kemampuan untuk menyediakan klasifikasi otomatis memiliki banyak keuntungan termasuk:

• Menandai hasil pencarian saat muncul

• Membatasi pencarian ke kategori tertentu (mengurangi kesalahan dan ambiguitas)

• Bantuan ke navigasi situs yang memungkinkan pengguna untuk menemukan bagian yang relevan dengan cepat

• Mengidentifikasi dokumen / produk / pengguna serupa sebagai bagian dari rekomendasi mesin (lihat di bawah)

• Meningkatkan komunikasi dan perencanaan dengan menyediakan bahasa yang sama (disebut sebagai ‘ontologi’)

Klasiﬁkasi digunakan secara luas dalam pencarian perusahaan seperti yang disediakan oleh Otonomi (http://www.autonomy.com/), dan banyak alat seperti filter spam dibangun berdasarkan prinsip klasi ﬁ kasi. Istilah representasi vektor dari suatu dokumen membuatnya relatif mudah dibandingkan dua atau lebih dokumen. Classi fi ers perlu diberikan seperangkat pelatihan contoh dokumen di mana kategori setiap dokumen diidentifikasi. Setelah klasifikasi memiliki membangun sebuah model berdasarkan dokumen pelatihan, model kemudian dapat digunakan untuk secara otomatis mengklasifikasikan dokumen baru. Generasi model biasanya dilakukan oleh algoritma pembelajaran mesin seperti naif Bayes, jaringan syaraf tiruan, vektor pendukung mesin (SVM) atau algoritma evolusi. Masing-masing metode berbeda memilikinya kekuatan dan kelemahan sendiri dan mungkin lebih berlaku untuk domain tertentu.

Metode pengkodean yang biasanya berbeda digabungkan.

**7.9 Pengukuran Kinerja Pengambilan**

Tujuan dari permintaan pencarian adalah untuk mengambil semua dokumen yang relevan dengan permintaan dan kembalikan tidak ada dokumen yang tidak relevan. Namun, untuk kumpulan data dunia nyata di atas ukuran dan kerumitan minimum, situasinya cenderung sama dengan yang ditunjukkan pada Gambar 7.3 di mana banyak dokumen yang relevan terlewatkan dan banyak dokumen diambil tidak relevan dengan kueri. Saat menguji mesin pencarian atau klasi ﬁ kasi, penting untuk dapat memberi

beberapa ukuran efektivitas. Ingat ukuran seberapa baik sistem pencarian menemukan dokumen yang relevan dan ketepatan mengukur seberapa baik sistem memilah-milah dokumen evant. Kami hanya dapat memperoleh nilai untuk penarikan dan presisi ketika satu set dokumen yang relevan dengan permintaan tertentu sudah diketahui. Ukuran F1 adalah cara yang umum digunakan untuk menggabungkan keduanya

langkah-langkah untuk memberikan angka efektivitas keseluruhan dan memiliki keuntungan memberi berat yang sama dengan presisi dan ingat. Akurasi sebenarnya yang dicapai akan bergantung pada sejumlah faktor seperti algoritma pembelajaran yang digunakan dan ukuran dan kualitas dokumen pelatihan. Kita harus mencatat bahwa keakuratan pada akhirnya turun ke manusia pertimbangan. Bahkan menggunakan klasifikasi manusia tidak akan mencapai akurasi 100% seperti pada sebagian besar domain dua manusia cenderung tidak setuju tentang label kategori untuk beberapa dokumen dalam koleksi ukuran dan kompleksitas yang wajar. Hasil yang mengesankan teknologi pembelajaran mesin terletak pada kenyataan bahwa dengan satu set dokumen contoh yang baik , telah dilaporkan bahwa otomatis klasifikasi dapat mencapai akurasi dekat bahwa para ahli manusia yang diukur menggunakan F1 atau serupa.

**7.10 Clustering**

Clustering adalah tugas lain yang bermanfaat, terutama ketika dilakukan secara otomatis. Di dalam Kasus tidak ada dokumen pelatihan berlabel yang diberikan, dan algoritma pengelompokan adalah diperlukan untuk menemukan 'kategori alami' dan mengelompokkan dokumen yang sesuai. Beberapa algoritma seperti k-means yang banyak digunakan mensyaratkan bahwa jumlah kategori disediakan di muka, tetapi sebaliknya tidak ada pengetahuan sebelumnya tentang koleksi diasumsikan. Sekali lagi, istilah vektor dokumen dibandingkan, dan kelompok dokumen dibuat berdasarkan kesamaan.

Dengan mengatur koleksi ke dalam kelompok, sejumlah tugas seperti penjelajahan dapat dilakukan menjadi lebih efisien. Seperti klasi ﬁ kasi, mengelompokkan deskripsi produk ucts / blogs / pengguna / dokumen atau data Web 2.0 lainnya dapat membantu meningkatkan pengguna pengalaman. Identifikasi otomatis barang-barang sejenis merupakan komponen penting dalam sejumlah alat lain seperti mesin rekomendasi (lihat di bawah).

Klasiﬁkasi disebut sebagai tugas 'pembelajaran yang diawasi' dan umumnya lebih akurat daripada pengelompokan yang 'tidak diawasi' dan tidak bermanfaat set dokumen pelatihan berlabel. Contoh pengelompokan yang baik adalah http: //search.yippy.com/ (dulu clusty.com). Buka situs dan cari ‘jawa’.

Contoh lain disediakan oleh Google News (http://news.google.com/) yang secara otomatis mengumpulkan ribuan artikel berita dan kemudian mengaturnya berdasarkan subjek, subtopik dan wilayah dan dapat dibuat untuk menampilkan hasil dengan cara yang dipersonalisasi. Umumnya, pengelompokan digunakan bersama dengan alat lain (misalnya klasiﬁkasi aturan) untuk menghasilkan hasil akhir.

**7.11 Penambangan Struktur Web**

Komponen mendasar dari web tentu saja adalah hyperlink. Link halaman cara bersama-sama dapat digunakan untuk lebih dari sekadar navigasi di browser, dan cukup banyak penelitian telah menyelidiki topologi hyperlink. Kita bisa mempertimbangkannya halaman web menjadi simpul pada grafik berarah (sederhanakan di mana garis di antara simpul-simpul pada grafik menunjukkan arah), dan di mana ada tautan dari halaman tertentu (p) ke halaman lain (q), kita dapat menganggap ini sebagai petunjuk. Jadi untuk halaman apa pun, kami dapat menetapkan nilai berdasarkan jumlah tautan yang mengarah ke dan dari halaman itu. Tingkat keluar dari simpul p adalah jumlah simpul yang dihubungkannya, dan di dalamnya derajat p adalah jumlah simpul yang memiliki tautan ke p. Dengan demikian, pada Gambar 7.4 halaman 1 memiliki tingkat di luar 1, dan halaman 2 memiliki derajat 2. Tautan dari halaman 2 ke halaman 3 dianggap lebih penting daripada tautan lain yang ditampilkan karena yang lebih tinggi di tingkat halaman 2. Dalam penambangan struktur web, kami melihat topologi web dan tautan di antaranya halaman bukan hanya pada isi halaman web. Hyperlink dari satu halaman ke yang lain dapat dianggap sebagai dukungan atau rekomendasi.

**7.11.1 HITS**

Pada tahun 1998 Kleinberg menciptakan algoritma HITS yang merupakan metode menggunakan tautan

struktur untuk menetapkan nilai penting ke halaman web yang dapat diperhitungkan peringkat hasil kueri. Algoritma memiliki beberapa kerugian termasuk fakta bahwa perhitungan dilakukan setelah permintaan diajukan waktu respons agak lambat.

**7.11.2 PageRank**

Pendiri Google (Larry Page dan Sergey Brin) membuat PageRank algoritma yang memberikan perkiraan pentingnya halaman. PageRank terlihat pada jumlah tautan yang menunjuk ke halaman web dan kepentingan relatif dari tautan tersebut dan karenanya tidak hanya tergantung pada jumlah tautan tetapi juga pada kualitas tautan. PageRank menegaskan bahwa jika halaman memiliki tautan penting yang menunjuk untuk itu, tautannya sendiri ke halaman lain juga penting. The PageRank computasi dapat dibuat sebelum pengguna memasukkan kueri sehingga waktu responsnyasangat cepat.

**7.13 Pencarian Multimedia**

Ada minat yang meningkat dan investasi besar dalam pengembangan baru teknologi untuk mencari konten audio dan visual. Dalam bentuknya yang paling sederhana, ini adalah sebuah kasus memberikan pertimbangan khusus pada teks dalam tag HTML untuk data multimedia. Namun, banyak vendor telah melampaui ini dan sekarang menyediakan 'deep ’search’ multimedia data termasuk fasilitas seperti mencari dengan gambar daripada teks. Kemampuan untuk secara otomatis mengelompokkan atau mengklasifikasikan data multimedia juga telah membuat kemajuan besar.

**7.14 Intelijen Kolektif**

Kecerdasan kolektif adalah bidang penelitian yang aktif tetapi bukan merupakan fenomena baru.

enon, dan contoh-contoh dikutip dalam biologi (misalnya proses evolusi), ilmu social dan ekonomi. Dalam bab ini, kami menggunakan kecerdasan kolektif untuk merujuk pada jenis kecerdasan yang muncul dari kolaborasi kelompok dalam web dan cloud-based sistem. Kecerdasan kolektif seperti ‘cloud’ adalah istilah di mana definisi yang tepat belum disetujui meskipun sebagian besar sudah jelas ketika kecerdasan kolektif sedang berlangsung dimanfaatkan. Pusat Kecerdasan Kolektif di Massachusetts Institute of Teknologi (http://cci.mit.edu/) memberikan deskripsi yang bagus dengan mengajukan pertanyaan alat-alat yang berdasarkan kecerdasan kolektif mungkin menjawab:

Bagaimana orang dan komputer dapat terhubung sehingga secara kolektif mereka bertindak lebih cerdas

daripada yang pernah dilakukan oleh individu, kelompok, atau komputer sebelumnya?

Munculnya web dan Web 2.0 telah menyebabkan sejumlah besar konten buatan pengguna. Pengguna tidak lagi membutuhkan keahlian khusus untuk menambahkan konten web mereka sendiri, dan memang Aplikasi Web 2.0 aktif mengundang pengguna untuk berinteraksi dan berkolaborasi. Cloud computing telah mempercepat proses sebagai kemudahan akses, ketersediaan penyimpanan, peningkatan efisiensi, skalabilitas dan biaya yang berkurang telah meningkatkan daya tarik penyimpanan informasi di web dan juga mempermudah dan mempermudah untuk inovasi baru aplikasi yang berusaha memanfaatkan kecerdasan kolektif untuk dikembangkan. Ada banyak contoh kecerdasan kolektif di mana sistem memanfaatkan kekuatan komunitas pengguna. Kami daftar beberapa di bawah ini untuk memberikan informasi tentang topik ini:

1. Wikipedia adalah hasil luar biasa dari kolaborasi besar, dan setiap artikel adalah dikelola secara efektif oleh banyak koleksi individu.

2. Banyak situs, seperti reddit.com, memungkinkan pengguna untuk menyediakan konten dan kemudian memutuskan melalui voting 'apa yang baik dan apa yang sampah'.

3. Google Suggest menyediakan daftar situs segera setelah pengguna mulai mengetik. Itu saran didasarkan pada sejumlah faktor, tetapi yang utama adalah yang sebelumnya pencarian yang dilakukan oleh individu (lihat Gambar 7.4).

4. Campuran Genius dari iTunes secara otomatis menyarankan lagu yang seharusnya berjalan dengan baik

dengan yang sudah ada di pustaka pengguna.

5. Facebook dan jejaring sosial lainnya menyediakan sistem untuk 'mencari teman' berdasarkan pada faktor-faktor seperti tautan yang sudah ada di antara teman-teman.

6. Toko online seperti Amazon akan memberikan rekomendasi berdasarkan sebelumnya pembelian.

7. Image tag bersama dengan perangkat lunak canggih seperti pengenalan wajah dapat digunakan untuk secara otomatis memberi label gambar tanpa tag.

8. Algoritma yang digunakan oleh situs penyewaan film Net fl ix untuk merekomendasikan films kepada pengguna sekarang bertanggung jawab atas 60% penyewaan dari situs ini.

9. Google + jaringan sosial memungkinkan pengguna untuk memberi +1 situs apa pun sebagai sarana rekomendasi.

10. Sistem email seperti Gmail akan menyarankan orang untuk memasukkan saat mengirim email (Gbr. 7.5)

Dalam bukunya, Wisdom of the Crowds, James Surowiecki, menunjukkan bahwa 'kelompok-kelompok itu

luar biasa cerdas, dan sering lebih pintar daripada orang terpintar di dalamnya '. Dia melanjutkan dengan berpendapat bahwa dalam keadaan yang tepat, hanya menambahkan lebih banyak orang akan mendapatkan hasil yang lebih baik yang seringkali lebih baik daripada yang dihasilkan oleh kecil sekelompok 'pakar'. Sejumlah ketentuan disarankan untuk ‘orang banyak yang bijaksana’:

1. Keanekaragaman: kerumunan terdiri dari individu dengan beragam pendapat.

2. Independensi: individu merasa bebas untuk mengekspresikan pendapat mereka.

3. Desentralisasi: orang dapat mengkhususkan diri.

4. Agregasi: ada beberapa mekanisme untuk mengumpulkan informasi itu dan menggunakannya untuk mendukung pengambilan keputusan.

Perhatikan bahwa kita tidak berbicara tentang membangun konsensus atau kompromi melainkan sebuah kompetisi antara pandangan heterogen; memang paralel tertentu telah tertarik dengan evolusi Darwin.

Algoritma PageRank yang dibahas di atas sebenarnya merupakan contoh kolektif kecerdasan sebagai pemesanan sebagian merupakan hasil dari suatu bentuk pemungutan suara yang terjadi dalam bentuk tautan antar situs web, dan kami dapat melihat bagaimana tautan tersebut dengan mudah memenuhi keempat kriteria yang ditetapkan atas. Web adalah lingkungan yang sangat beragam, dan individu independen bebas untuk menambahkan tautan ke halaman lain dari halaman mereka sendiri. Sistem yang mendasarinya sudah terkenal desentralisasi dan agregasi adalah pusat dari algoritma PageRank. Pencarian sebenarnya mesin sebenarnya menggunakan campuran kompleks algoritma berbasis matematis, penambangan teks dan teknik pembelajaran mesin bersama dengan sesuatu yang mirip dengan PageRank. Ayo lihat lagi beberapa contoh ‘Inteligensi Kolektif yang Beraksi’.

**7.14.1 Pemberian tag**

Pemberian tag adalah suatu proses yang mirip dengan klasi ﬁ kasi di mana barang-barang seperti produk, web halaman dan dokumen diberi label. Dalam kasus klasifikasi teks otomatis, kami melihat bahwa kedua kategori kategori yang ditetapkan dan satu set dokumen contoh yang diberi label diminta untuk membangun klasifikasi. Dalam hal ini orang dengan beberapa keahlian dalam domain akan diminta untuk memberikan setiap contoh dokumen label kategori yang sesuai. Salah satu pendekatan untuk memberi tag adalah menggunakan profesional untuk memberi label item, meskipun ini bisa mengambil sumber daya yang signifikan dan bisa menjadi tidak layak, misalnya, di mana ada sejumlah besar konten buatan pengguna. Pendekatan kedua adalah menghasilkan tag secara otomatis baik melalui pengelompokan atau

klasiﬁkasi sebagaimana dijelaskan di atas atau dengan menganalisis teks. Sekali lagi kita bisa menggunakan istilah itu vektor untuk mendapatkan bobot relatif dari istilah yang ditemukan dalam teks yang dapat dikirim ke pengguna dalam berbagai format visual (lihat di bawah). Opsi ketiga, yang memiliki

terbukti semakin populer dalam beberapa tahun terakhir dan memang telah ada di mana-mana di web, adalah untuk memungkinkan pengguna membuat tag mereka sendiri, baik dengan memilih tag dari daftar predikat atau dengan menggunakan kata atau frasa apa pun yang mereka pilih untuk melabeli item. Pendekatan ini, terkadang disebut sebagai ‘folksonomy’ (kombinasi dari ‘folk’dan 'taksonomi'), bisa jauh lebih murah daripada mengembangkan taksonomi yang dikontrol dan memungkinkan pengguna untuk memilih istilah yang mereka temukan yang paling relevan. Dalam sebuah folksonomy pengguna menciptakan klasifikasi mereka sendiri, dan ketika mereka melakukan kecerdasan kolektif alat dapat mengumpulkan informasi tentang barang yang ditandai dan tentang pengguna yang sedang membuat tag. Setelah sejumlah item yang wajar telah ditandai, kami bias kemudian gunakan barang-barang tersebut sebagai contoh untuk klasi ﬁ kasi yang kemudian dapat dicari secara otomatis barang serupa. Tag dapat membantu dengan menemukan pengguna dengan item yang ditandai serupa, menemukan item yang mirip dengan yang diberi tag dan membuat dinamis, berpusat pada pengguna kosakata untuk analisis dan kolaborasi dalam domain tertentu.

**7.14.2 Mesin Rekomendasi**

Mesin rekomendasi adalah contoh utama dari 'Kolektif Intelijen Beraksi'. Pada tahun 2006, Netflix mengadakan Penghargaan netflix pertama, menemukan program untuk memprediksi preferensi lebih baik preferensi pengguna dan mengalahkan sistem rekomendasi film Netflix yang ada paling sedikit 10%. Itu Hadiah $ 1 juta dimenangkan pada tahun 2009, dan mesin rekomendasi sekarang dilaporkan untuk berkontribusi pada sebagian besar dari keseluruhan proyek. Mungkin rekomendasi terbaik yang diketahui

mesin pendation berasal dari Amazon. Jika kita memilih buku 'Lucene in Action', kami dapatkan layar yang ditunjukkan pada Gambar 7.6 di mana dua set saran ditampilkan. Bahkan jika Anda mengklik pada beberapa saran Anda akan segera mengambil sebagian besar Bacaan yang direkomendasikan untuk bab ini. Seperti yang ditunjukkan pada halaman, informasinya sebagian besar diperoleh dengan menganalisis pola pembelian pelanggan sebelumnya. Banyak fitur lain dapat dimasukkan ke dalam rekomendasi seperti kesamaan item dianalisis, kesamaan antara pengguna dan kesamaan antara pengguna dan item.

**7.14.3 Intelijen Kolektif di Perusahaan**

Dalam organisasi besar, karyawan dapat membuang waktu yang signifikan dalam mencoba mencari dokumen dari pencarian mereka. Perusahaan pencarian perusahaan seperti Recommind (http: //www.recommind.com/) mendorong pengguna untuk memberi tag dan menilai item. Informasinya

dikumpulkan dan kemudian dimasukkan ke mesin pencari sehingga pengguna lain dapat dengan cepat mendapatkan merasa untuk topik tertentu dari tag dan kegunaan yang berasal dari klasiﬁkasi dan rekomendasi orang lain. Ini juga memungkinkan untuk opsi menghapus atau mengurangi peringkat barang yang dinilai buruk, suatu proses yang kadang-kadang disebut sebagai penyaringan kolaboratif.

**7.14.4 Penilaian Pengguna**

Banyak situs e-commerce memberi pengguna opsi untuk mendaftarkan pendapat mereka tentang suatu produk, seringkali dalam bentuk sistem bintang sederhana. Peringkat pengguna dapat dikumpulkan dan dilaporkan kepada calon pembeli. Seringkali pengguna juga dapat menulis ulasan yang dapat dibaca oleh pelanggan lain dan banyak situs kini menawarkan tautan ke situs jejaring sosial seperti Facebook dan Twitter tempat produk dapat didiskusikan. Peringkat pengguna tertentu juga dapat dievaluasi dengan membandingkan dengan yang diberikan oleh pengguna lain. Baru-baru ini ada minat yang signifikan dalam alat-alat yang dapat secara otomatis menambang ulasan tekstual dan identifikasi sentimen atau pendapat yang diungkapkan, khususnya dalam kasus data jejaring sosial. Perangkat lunak mungkin tidak sepenuhnya akurat karena tantangan teks bahasa alami yang disebutkan di atas tetapi sebagai tema berulang dalam kecerdasan kolektif, jika keakuratannya masuk akal dan jumlah ulasan yang ditambang di atas ambang batas minimal mungkin masih sangat bermanfaat dalam menggunakan alat ini.

**7.14.5 Personalisasi**

Layanan personalisasi dapat membantu dalam mengatasi informasi yang berlebihan. Informasi yang disajikan kepada pengguna dapat secara otomatis disaring dan diadaptasi tanpa intervensi eksplisit dari pengguna. Personalisasi biasanya didasarkan pada profil pengguna yang mungkin sebagian atau seluruhnya dihasilkan mesin. Baik struktur dan konten situs web dapat secara dinamis disesuaikan dengan pengguna berdasarkan profil mereka. Sebuah kesuksesan sistem harus dapat mengidentifikasi pengguna, mengumpulkan pengetahuan tentang preferensi mereka dan kemudian membuat personalisasi yang sesuai. Tentu saja, tingkat kepercayaan pada sistem diperlukan, dan akan jauh lebih baik untuk melewati fungsi personalisasi daripada membuat perubahan berdasarkan asumsi yang salah yang mengarah ke merugikan pengalaman pengguna dan mengurangi efektivitas situs. Kami telah melihat penambangan konten web, setidaknya dalam hal data tekstual dan penambangan struktur web dalam hal PageRank. Pengguna juga memberikan informasi berharga melalui interaksi dengan situs web. Terkadang informasinya cukup eksplisit seperti dalam kasus pembelian, penilaian, penunjuk atau pemungutan suara. Dalam e-commerce, produk dan layanan dapat direkomendasikan kepada pengguna berdasarkan tidak hanya pada tindakan pembelian pengguna lain tetapi tergantung pada kebutuhan dan preferensi spesifik terkait dengan profil. Penambangan penggunaan web yang berhubungan dengan data yang dikumpulkan dari kunjungan pengguna ke situs sangat berguna untuk mengembangkan sistem personalisasi. Data seperti frekuensi kunjungan, halaman yang dikunjungi, waktu yang dihabiskan di setiap halaman dan rute yang dinavigasi melalui situs dapat digunakan untuk penemuan pola, sering dengan bantuan algoritma pembelajaran mesin, yang kemudian dapat dimasukkan ke dalam mesin personalisasi.

**7.14.6 Penonton Kerumunan**

Banyak sistem untuk memanfaatkan kecerdasan kolektif memerlukan sedikit atau tidak sama sekali

masukan tambahan dari pengguna. Namun, langsung bertanya kepada pelanggan atau tertarik pihak-pihak yang membantu proyek telah terbukti berhasil di sejumlah bidang. Salah satunya contoh paling awal dan paling terkenal berasal dari tahun 2000 ketika NASA memulai Penelitian ClickWorker yang menggunakan sukarelawan publik untuk membantu mengidentifikasi kawah. Ada sejumlah sistem yang bermunculan seperti IdeaExchange dari Salesforce.com (http://success.salesforce.com/) tempat pelanggan dapat mengusulkan yang baru solusi produk dan memberikan evaluasi produk yang ada. Amazon Mechanical Turk (https://www.mturk.com/mturk/welcome) menggunakan crowdsourcing untuk menyelesaikan berbagai variasi masalah yang membutuhkan kecerdasan manusia.

**7.15 Visualisasi Teks**

Kami menyelesaikan bab ini dengan pengantar singkat ke bidang visualisasi teks. Sudah lama diketahui bahwa di mana ada sejumlah besar informasi untuk dicerna, visualisasi dapat menjadi bantuan besar bagi pengguna data tersebut. Beberapa tahun terakhir telah melihat seperangkat alat yang berkembang untuk secara otomatis memvisualisasikan konten teks dokumen dan halaman web. Tag cloud adalah visualisasi sederhana yang menampilkan informasi frekuensi kata melalui ukuran font dan warna yang telah digunakan di web sejak 1997. Pengguna telah menemukan visualisasi yang berguna dalam memberikan ikhtisar konteks dokumen teks dan situs web. Sementara banyak sistem dibentuk menggunakan tag yang disediakan pengguna, ada minat yang signifikan dalam ‘kata cloud’ atau ‘tag teks’ yang dihasilkan secara otomatis menggunakan teks yang ditemukan di dokumen atau situs web. Misalnya, alat populer Wordle telah melihat peningkatan yang stabil dalam penggunaan, dan Wordle atau diagram serupa umumnya dilihat sebagai bagian dari situs web untuk memberikan pengguna pandangan cepat tentang topik-topik penting. Umumnya kata awan didasarkan pada informasi frekuensi setelah penghapusan kata berhenti dan mungkin berasal. Jika stemming digunakan, penting untuk menampilkan kata yang dapat dikenali (sering kali bentuk yang paling sering muncul) daripada bentuk stem yang mungkin membingungkan bagi pengguna. Kita dapat menganggap diagram sebagai representasi visual vektor istilah untuk dokumen dan seperti dalam representasi vektor istilah, itu didasarkan pada 'kantong kata-kata', dan kata kedekatan umumnya tidak diperhitungkan saat menghasilkan kata awan . Jika Anda pergi ke situs Wordle (http://www.wordle.net/), Anda dapat melihat contoh atau Anda dapat membuat awan kata Anda sendiri dengan menempelkan teks atau menunjuk ke sumber yang berisi teks. Sistem memilih kata yang paling sering dan kemudian menyajikannya menggunakan berbagai teknik untuk menyesuaikan font, warna, ukuran dan posisi, dengan cara yang menyenangkan dan bermanfaat bagi pengguna. Alternatif eksperimental tersedia di http://txt2vz.appspot.com/ Txt2vz mirip dengan Wordle tetapi juga menunjukkan bagaimana kata-kata dihubungkan dalam teks dan memiliki beberapa fitur dari peta pikiran. Dua format dari kata txt2vz cloud tersedia dan contoh-contoh yang dihasilkan dari teks bab ini ditunjukkan di bawah ini (Gambar. 7.7dan 7.8).

Awan kata sederhana dan biasanya disajikan di situs web dengan sedikit atau tanpa penjelasan tentang bagaimana mereka harus digunakan atau ditafsirkan. Seringkali kata-kata yang disajikan dibuat dapat diklik sebagai sarana untuk menyediakan alat navigasi dinamis yang menyesuaikan secara real time saat pengguna menambahkan konten.

**7.16 Ringkasan Bab**

Dalam bab ini, kami telah mengambil ikhtisar tentang berbagai cara di mana kita bisa mendapatkan hasil maksimal dari sistem berbasis web dan cloud. Secara khusus kami telah melihat aplikasi Web 2.0 dan kemungkinan menarik menggunakan alat cerdas untuk mendapatkan informasi berharga dan untuk meningkatkan pengalaman pengguna.