Indexing dan Bahasa Penelusuran

Sugeng Priyanto

Indexing

- Definisi : sebuah proses untuk melakukan pengindeksan terhadap kumpulan dokumen yang akan disediakan sebagai informasi kepada pemakai.
- Proses pengindeksan bisa secara manual ataupun secara otomatis.

Indeks

- Definisi : cantuman dari bermacam-macam atribut yang diharapkan dapat digunakan sebagai dasar pencarian dokumen.
- Jika atribut tersebut berupa subjek, maka indeks yang mewakilinya disebut sebagai indeks subjek. Sedangkan bila atribut tersebut berupa pengarang, maka indeks yang mewakilinya disebut sebagai indeks pengarang.
- untuk memungkinkan ditemukannya dokumen atau corpus (wakil dokumen) yang relevan dengan pertanyaan (query) dengan tepat.

Tahapan

- **Parsing** dokumen yaitu proses pengambilan kata-kata dari kumpulan dokumen.
- **Stoplist** yaitu proses pembuangan kata buang seperti: tetapi, yaitu, sedangkan, dan sebagainya.
- **Stemming** yaitu proses penghilangan/ pemotongan dari suatu kata menjadi bentuk dasar. Kata “diadaptasikan” atau “beradaptasi” mejadi kata “adaptasi” sebagai istilah.
- **Term Weighting** dan **Inverted File** yaitu proses pemberian bobot pada istilah.

Bahasa Penelusuran

- **Bahasa Alami (natural languanges)**
- **Kosa kata terkontrol (controlled vocabulary)**

Bahasa Alami (Natural languanges)

- Disebut juga dengan uncontrol vocabulary
- Definisi : bahasa dari dokumen yang di indeks biasanya bahasa tersebut merupakan bahasa yang tidak terkendali (uncontrolled vocabullary)

Kelebihan bahasa alami

- **Bahasa alami dapat dengan mudah dimengerti oleh pengguna tanpa harus memerlukan pelatihan khusus, dan berbagai nuansa makna dapat direpresentasikan dengan lebih leluasa.**
- **Bahasa alami memiliki spesifikasi yang tinggi. Spesifikasi istilah ini muncul karena dapat menggunakan seluruh istilah yang terdapat dalam setiap judul dan subjek sebagai query. Spesifikasi istilah akan memudahkan pencarian untuk mendapatkan ketepatan (precision) yang tinggi.**

- Bahasa alamiah memiliki kedalaman yang tinggi, artinya banyak tema atau subjek baru yang dihasilkan dokumen yang dapat dijadikan sebagai istilah baru dalam penelusuran.
- Penelusuran yang merupakan praktisi dalam bidangnya dapat melakukan penelusuran dengan bahasa alamiah yang dengan lebih selektif.

Kekurangan bahasa alamiah

- **Bahasa alamiah kurang ringkas. Query yang digunakan penelusuran sering berupa kata atau istilah tidak standar sehingga sering terjadi kehilangan informasi saat penelusuran**
- **Mempunyai ambiguitas yang tinggi. Ambiguitas adalah kata atau istilah yang dapat memiliki lebih dari satu arti sehingga mengakibatkan kerancuan. Ambiguitas dapat terjadi karena sinonim atau homograf.**
- **Ketidak mampuan komputer untuk menyerap atau menangkap makna dari suatu pernyataan.**

Keunggulan Kosa Kata Terkendali

- **Proses penelusuran dan temu balik informasi lebih efisien, Artinya, dengan menggunakan kosa kata terkendali seperti indeks subjek atau tesaurus dalam penelusuran, maka ketepatan dari dokumen yang terambil dengan kebutuhan pengguna dapat diperoleh dalam waktu yang relatif singkat.**

- Mempunyai representasi dokumen yang konsisten. Kosa kata atau istilah yang digunakan dalam pengindeksan dokumen pada saat input ke sistem adalah kosa kata yang terkendali dan standar. Oleh karena itu, bila kosa kata atau istilah tersebut kemudian dijadikan sebagai query untuk pencarian atau penelusuran, maka sudah pasti akan tetap mewakili atau merepresentasikan dokumen yang sama seperti pada saat input sistem dilakukan.

- Memudahkan penelusuran komprehensif dengan menyatukan istilah terkait secara semantis. Maksudnya, ada kalanya suatu kosa kata atau indeks subjek tertentu mempunyai hubungan makna dengan indeks yang lain, sehingga dapat digunakan untuk memperkuat pencarian.
- Memiliki ambiguity yang sangat kecil. Ambiguitas atau kerancuan dapat dikurangi sekecil mungkin karena kosa kata dapat mengontrol sinonim dan homograf.

Kelemahan

- Kosa kata terkendali harus selalu diperbaharui.
- Kosa kata terkendali (controlled vocabulary) sering dihadapkan kepada ketidak-cocokan (incompatibility) istilah di antara satu database dengan database lainnya pada bidang ilmu yang sama
- Kurangnya spesifikasi dalam kosa kata.
- Kosa kata terkendali memiliki struktur yang tidak lengkap.
- Kosa kata terkendali memerlukan biaya dan upaya yang besar pada waktu input sistem yaitu pada saat pengindeksan dilakukan

Pendekatan manakah yang ideal ?

Pertimbangan pemilihan

- Pencari informasi memerlukan spesifikasi dan ekspresi dalam merepresentasikan konsep pencarian.
- Istilah yang diperlukan untuk merepresentasikan konsep tidak terdapat pada bahasa terkendali
- Pencari informasi menginginkan pencarian topik yang menyeluruh termasuk hal-hal yang tidak berhubungan langsung dengan topik yang dimaksud.
- Subjek yang dicari bukan termasuk koleksi inti dari pangkalan data yang dipergunakan.