TAKSONOMI MODEL-MODEL SISTEM TEMU KEMBALI INFORMASI

MATERI PERKULIAHAN : *INFORMATION RETRIEVAL SYSTEM* KE-4

Disusun Oleh :
Nama : Nuning Kurniasih, S.Sos., M.Hum.
NIP. 197606252000122001

Departemen Ilmu Informasi dan Perpustakaan
Fakultas Ilmu Komunikasi
Universitas Padjadjaran
Ditulis Pertama Tahun 2005, Revisi Januari 2014

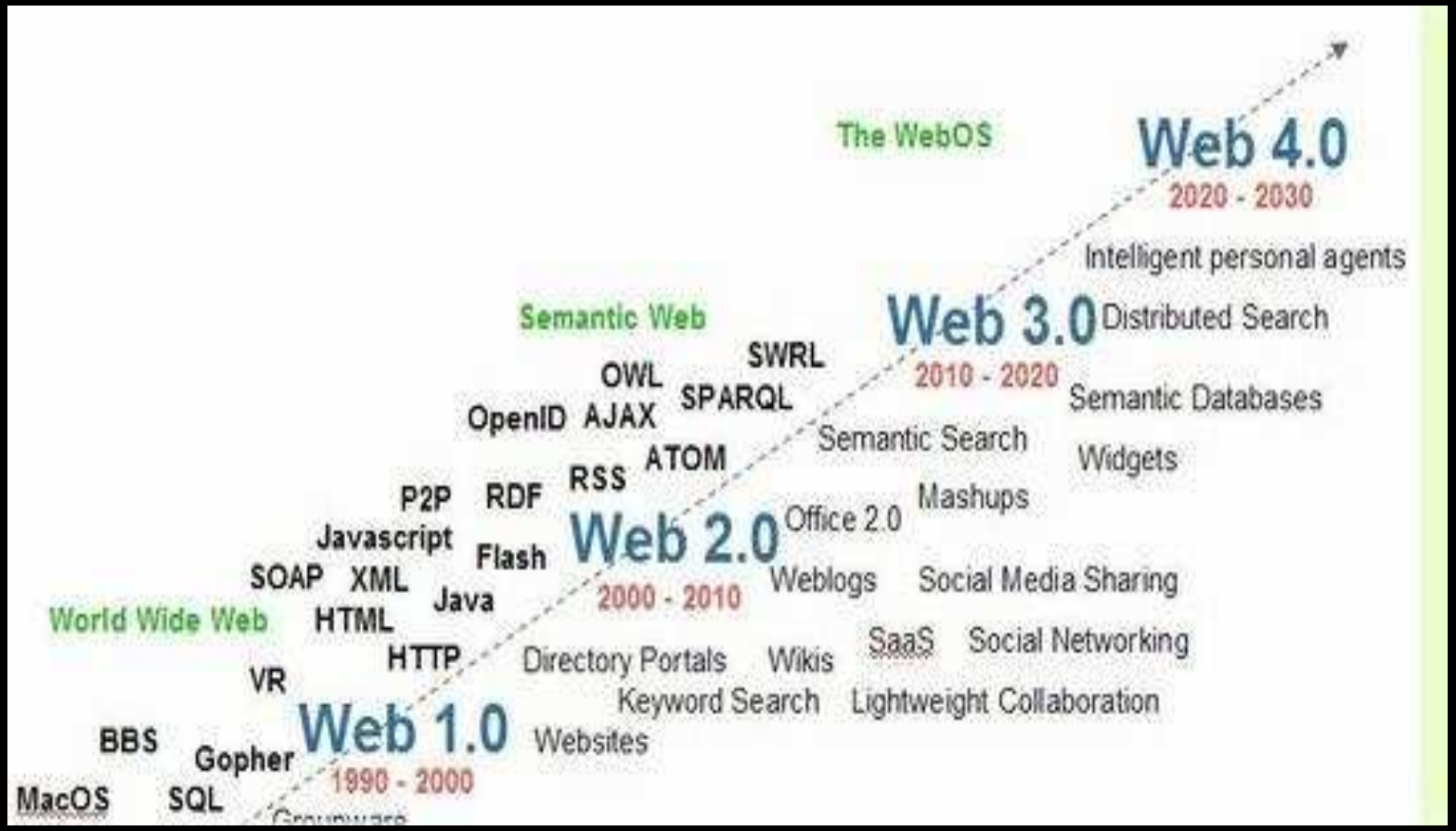
JENIS SISTEM TEMU KEMBALI INFORMASI

1. Lokal

2. Global

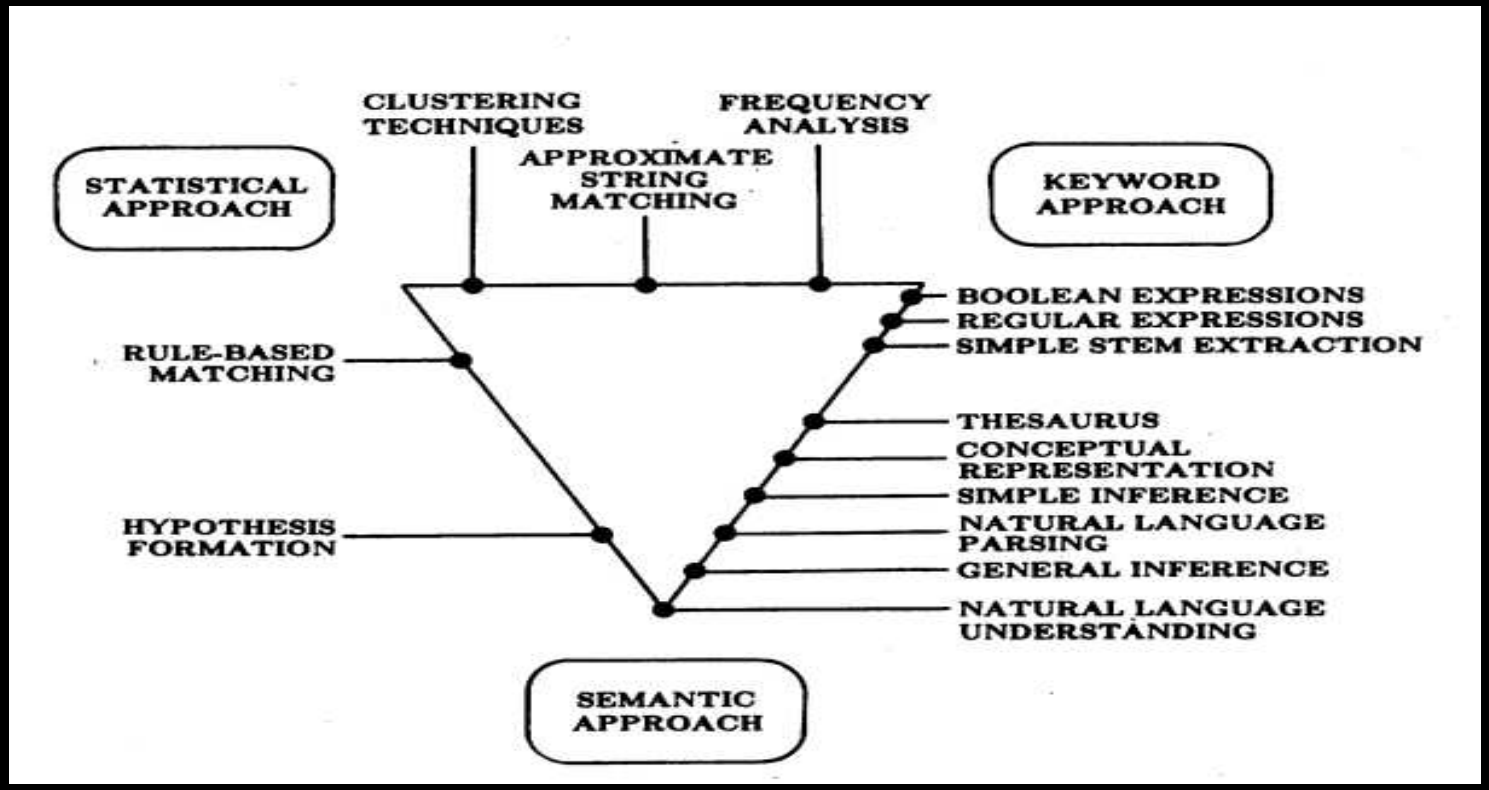
Baca kembali materi perkuliahan ke-2

World Wide Web



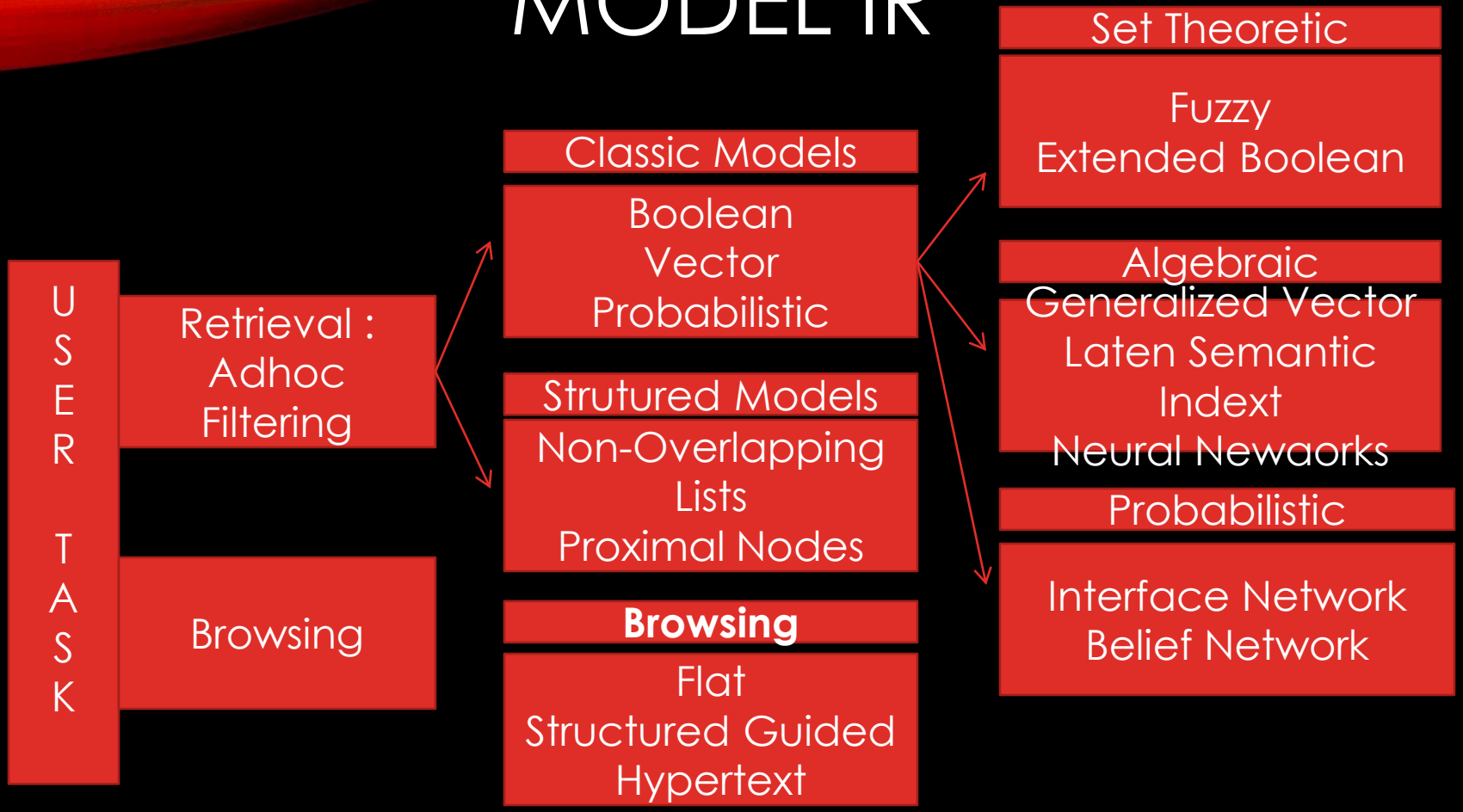
http://blog.law.cornell.edu

THE INFORMATION RETRIEVAL TRIANGLE



Mc Cune, Brian P. 1985.

MODEL IR



Baeza-Yates & Ribeiro Neto

LOGICAL VIEW OF DOC AND RETRIEVAL TASK

LOGICAL VIEW OF DOCUMENTS

USER TASK

	Index Term	Full Text	Full Text + Structure
Retrieval	Classic Set Theoretic Algebraic Probabilistic	Classic Set Theoretic Algebraic Probabilistic	Structured
Browsing	Flat	Flat Hypertext	Structured Guide Hypertext

Baeza-Yates & Ribeiro Neto

KONSEP DASAR MODEL-MODEL IR KLASIK

- Dokumen direpresentasikan dengan seperangkat *keyword* atau *index term* yang representative.
- Sebuah item indeks adalah kata-kata dalam dokumen yang mudah diingat untuk tema umum sebuah dokumen.
- *Index term* biasanya kata benda.
- *Search engines mengasumsikan semua kata adalah* index term (merepresentasikan full text)
- Tidak semua *term* dapat merepserentasikan isi dokumen, sebagian *term* dapat mengidentifikasi dokumen dengan sempit.

KONSEP DASAR MODEL-MODEL IR KLASIK

- *Kj* adalah *index term*
- *dj* adalah sebuah dokumen
- *t* adalah total dari jumlah dokumen
- *K = (k1, k2, …, kt)* adalah seluruh *index term*
- *wij>=0* is penimbang yang berasosiasi dengan (ki, dj)
- *wij=0* mengindikasikan bahwa term tidak berhubungan dengan dokumen
- *vec (dj)=(w1j, w2j,…, wtj)* adalah penimbang *vector* yang berasosiaasi dengan dokumen *dj*.
- *gi(vec(dj))=wij* adalah sebuah fungsi dengan pengembalian pertimbangan yang berasisiasi dengan pasangannya associated with pair *(kj, dj)*

MODEL TEMU KEMBALI INFORMASI KLASIK

- Boolean Query
  - Operator Bolean Logic : AND, OR, NOT
- Vector Query
  - Seberapa mirip dokumen dengan?
    - [(java 3) (compiler 2) (unix 1) (linus 1)]
- Probabilistic Query
  - Probalistik Relevan – Pr(rel)
  - Probabilistik Non Relevan – Pr(non rel)

MODEL PENELUSURAN BOOLEAN

Temu kembali berbasis pada kriteria pengambilan keutusan dua kondisi (binary) tanpa ide atau pencocok yang pasrial.

Tidak ada dalam dokumen yang disediakan.

Informasi perlu diterjemahkan kedalam sebuah *Boolean expression*.

Operator *Boolean Logic* diekspresikan dengan

Logical AND

Logical OR

Logical NOT

BOOLEAN STANDAR

	Standard Boolean
Goal	• Capture conceptual structure and contextual information
Methods	• Coordination: AND, OR, NOT • Proximity • Fields • Stemming / Truncation
(+)	• Easy to implement • Computationally efficient => all the major on-line databases use it • Expressiveness and Clarity Synonym specifications (OR-clauses) and phrases (AND-clauses).
(-)	• Difficult to construct Boolean queries. • All or nothing AND too severe, and OR does not differentiate enough. • Difficult to control output: Null output <-> Overload. • No ranking • No weighting of index or query terms • No uncertainty measure

Ringkasan dari Karakteristik Standar Pendekatan Boolean disertai Kelebihan dan Kelemahannya
http://comminfo.rutgers.edu

LOGICAL AND

- Memperbolehkan penelusur untuk menggunakan pernyataan query ke dalam sua tau lebih konsep sehingga hasil penelusuran menjadi lebih terbatas.
- Formula penyataaan sederhana A AND B
- Contoh untuk menelusur *marketing* ***and*** *library*, kita memformulasikan penyataan penelusuran dengan :

  marketing ***AND*** *library*

- Dengan query tersebut maka kita akan menemukan dokumen yang mengandung unsur marketing dan perpustakaan saja, dan tidak untuk mendapatkan dokumen yang hanya mengandung unsur marketing atau perpustakaan saja.

LOGICAL OR

- Memperbolehkan penelusur untuk secara spesifik menggunakan alternatif *term* (atau konsep) yang mengindikasikan dua konsep sesuai dengan tujuan penelusuran. Hal ini menjadikan hasil penelusuran menjadi lebih luas, karena adanya alternatif dalam penyataan query.
- Formulasi pernyataan sederhana : A OR B.
- Contoh :

  marketing OR library

Dengan query tersebut maka kita akan mendapatkan dokumen yang mengandung unsur marketing saja, perpustakaan saja, atau yang mengandung unsur marketing dan perpustakaan.

LOGICAL NOT

- Dapat mengecualikan item-item dari seperangkat term penelusuran.
- Pernyataan formulasi sederhana :

  A NOT B

- Contoh :

  marketing NOT library

  Ini artinya kita hanya menginginkan dokumen yang mengandung unsur marketing yang di dalamnya tidak ada unsur perpustakaannya..

KOMBINASI LOGICAL AND, OR, NOT

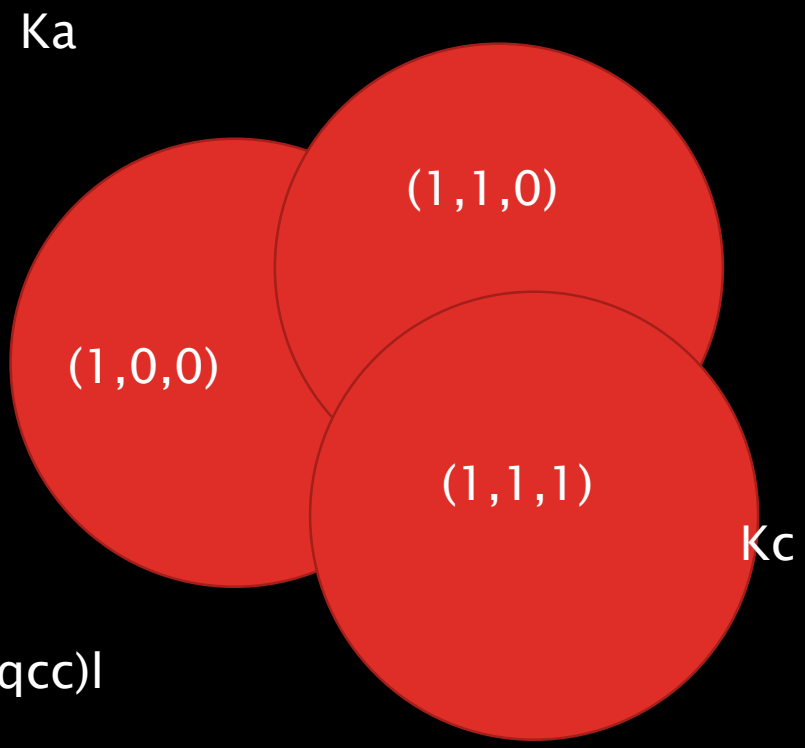
- Dapat mengkombinasikan satu pernyataan ke dalam penelusuran yang kompleks.
- Contoh :

  marketing AND library OR information centre NOT profit organization

Artinya kita ingin mendapatkan dokumen yang mengandung untur marketing dan perpustakaan tanpa unsur pusat informasi dan bukan untuk organisasi non profit.

DIAGRAM VEN UNTUK MODEL BOOLEAN

q = ka ^(kbv¬kc)

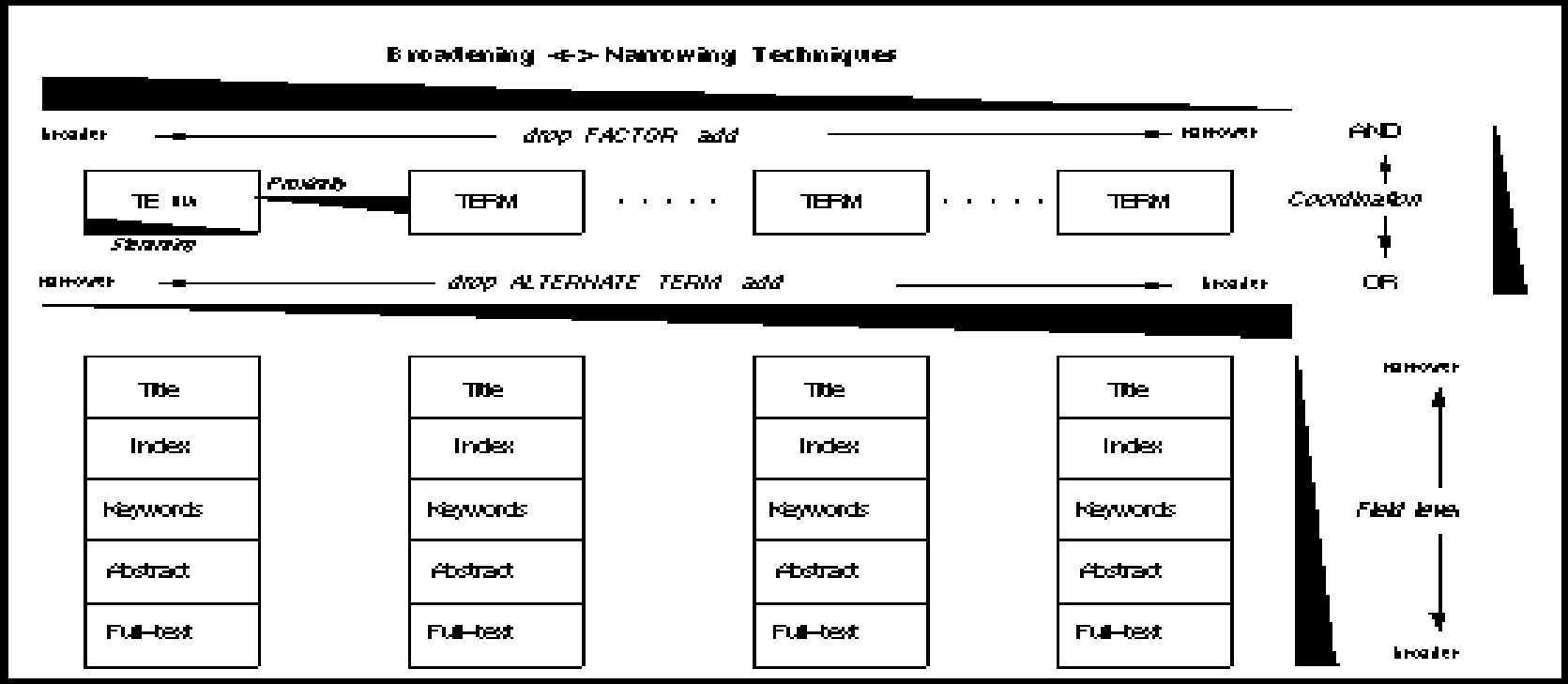


Sim (q,dj)=1 if Ǝ vec (qcc)l

(vec(qcc)ϵvec(qdnf))^
(Vki,
gi(vec(dj))=gi(vec(qcc)))
0 otherwise

Baeza-Yates & Ribeiro Neto

TEKNIK MEMPERLUAS DAN MEMPERSEMPIT



Teknik Memperluas dan Mempersempit Hasil Penenlusuran Menggunakan Boolean Logic
http://comminfo.rutgers.edu

SMART BOOLEAN

	Smart Boolean
Goal	• Structure search (re-)formulation process. • Use structural and contextual knowledge-bases and clarity of Boolean expressions.
Methods	• Natural language statement is automatically translated into Boolean Topic Representation • Boolean Topic Representation: ANDs of ORs of concepts → Keyword/stem, all fields • Conceptual info. → Coordination and Add/Drop Factor • Contextual info. → Proximity • Structural info. → Field levels • Synonym or word relationships → Stemming/Truncation overlap => all this information can be used to rank documents • Techniques to Broaden and Narrow query
(+)	• No need for Boolean operators => Convert operator-free statement into ANDs of ORs • Assist user in query (re)formulation: by asking users targeted questions to automatically modify the query. • "Why irrelevant?" -> activates narrowing methods. • "Broaden by Dropping Factors" to estimate recall.
(-)	• How to visualize ? • Conceptual query representation (BTR) • Query modification techniques and their effects • Structured relevance feedback

http://comminfo.rutgers.edu

MODEL EXTENDED BOOLEAN

	Extended Boolean Models
Goal	• Less strict Boolean operators • Ranked output
Methods	• **Fuzzy logic** [OR -> max], [AND -> min] and [NOT -> 1 – max] (–) Lack of sensitivity of min and max: min(0.2, 0.8) = min(0.2, 0.3).

Model Perluasan Boolean digunakan untuk :

1. Apabila operator Boolean terlalu ketat dan perlu diperhalus.
2. Pendekatan standar Boolean tidak menyediakan fitur ranking sehinggan pendekatan dan metode pendeskripsian dapat membantu perankingan dokumen yang relevan [Fox and Koll 1988, Marcus 1991].
3. Model Boolean tidak mendukung tugas pertimbangan query atau term dari sebuah dokumen.

http://comminfo.rutgers.edu

VECTOR DAN PROBABILISTIK

Statistical	Vector Space	Probabilistic
Motivation	Simplify query formulation Ability to control output	Address uncertainty in query representations
Goal	Rank the output based on Similarity	Probability of Relevance
Methods	Cosine measure	Use of different models
Source	**Query Term Statistics** Vector-Space: • similarity(Q,D) $= \Sigma (w_{iq} \times w_{ij}) / \text{"normalizer"}$ where $w_{iq} = (0.5 + 0.5\ freq_{iq} / maxfreq_q) \times \mathbf{idf}(i)$ $w_{ij} = freq_{ij} \times \mathbf{idf}(i)$ • inverse term freq. in collection $\mathbf{idf}(i) = \log_2 (N - n(i)) / n(i)$. Probabilistic: • term weight $= \log [(r_t / R - r_t) / ((n_t - r_t) / ((N - n_t) - (R - r_t)))]$ $=$ "(hits / misses) / (false alarms / correct misses)" • similarity $_{jk} = \Sigma (C + \mathbf{idf}(i)) \times \mathbf{tf}(i,j)$ where $\mathbf{tf}(i,j) = K + (1-K)\ (freq(i,j) / maxfreq(j))$.	
Issues	• How to express NOT ? Proximity searches ? • Limited expressive power • Computationally intensive • Assumes that terms are independent. • Lack of structure to represent important linguistic features • How to better visualize the retrieved set ?	• Estimation of needed probabilities • Prior knowledge needed. • Independence assumption • Boolean relations lost. • Which model is best ?

http://comminfo.rutgers.edu

VECTOR QUERY

- Koleksi dokumen *n* dengan perbedaan term *t* dapat direpresentasikan dengan sebuah matrix.

	T_1	T_2	....	T_t
D_1	w_{11}	w_{21}	...	w_{t1}
D_2	w_{12}	w_{22}	...	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	...	w_{tn}

- Sebuah query juga dapat direpresentasikan sebagai sebuah vector seperti sebuag dokumen.

LATIHAN

- Temu kembali informasi dalam sebuah database dengan menggunakan operator boolean logic operators.

Contoh, kunjungin http://online.sagepub.com dan telusur beberapa tema tugas perkuliahan.

Terima Kasih

Contact Me @nuningkurniasih