CS-463 Information Retrieval Systems

Μοντέλα Ανάκτησης (Retrieval Models)

Part C

Yannis Tzitzikas

University of Crete

CS-463,Spring 05

Lecture : 5
Date : 8-3-2005

Διάρθρωση Διάλεξης

PART (A)

- *Ανάκτηση και Φιλτράρισμα*
- *Εισαγωγή στα Μοντέλα Άντλησης*
- *Κατηγορίες Μοντέλων*
- *Exact vs Best Match*
- *Τα κλασσικά μοντέλα ανάκτησης*
  - *Το Boolean Μοντέλο*
  - *Στατιστικά Μοντέλα - Βάρυνση Όρων*
  - *Το Διανυσματικό Μοντέλο*
  - *Το Πιθανοκρατικό Μοντέλο*

PART (B):***Εναλλακτικά μοντέλα***

- *(I) Συνολοθεωρητικά μοντέλα*
  - *Fuzzy Retrieval Model*
  - *Extended Boolean Model*
- *(II) Αλγεβρικά Μοντέλα*
  - *Latent Semantic Indexing*
  - *Neural Netwok Model*

PART (C):

- ***(III) Πιθανοκρατικά Μοντέλα***
  - *Bayesian Network Model*
  - *Inference Network Model*

Εναλλακτικά Πιθανοκρατικά Μοντέλα

- Probability Theory
  - Semantically clear
  - Computationally clumsy
- Why Bayesian (Belief) Networks?
  - Clean formalism to combine distinct sources of evidence
    - past queries, past feedback cycles, distinct query formulations
  - Modularize the world (dependencies)
- Bayesian Network Models for IR:
  - Inference Network (Turtle & Croft, 1991)
  - Belief Network (Ribeiro-Neto & Muntz, 1996)

Bayesian Inference

Basic Probability Axioms:

$0 \leq P(A) \leq 1$;

P(sure)=1;

P(A V B)=P(A)+P(B) if A and B are mutually exclusive

Other formulations

- $P(A)=P(A \wedge B)+P(A \wedge \neg B)$
- $P(A)= \Sigma_{\forall i} P(A \wedge B_i)$, where $B_{i,\forall i}$ is a set of exhaustive and mutually exclusive events
- $P(A) + P(\neg A) = 1$
- $P(A|K)$ belief in A given the knowledge K
- if P(A|B)=P(A), we say:A and B are *independent*
- *if* $P(A|B \wedge C)= P(A|C)$, we say: A and B are conditionally independent, given C
- $P(A \wedge B)=P(A|B)P(B)$
- $P(A)= \Sigma_{\forall i} P(A \mid B_i)P(B_i)$

Bayesian Inference

Bayes' Rule : the heart of Bayesian techniques

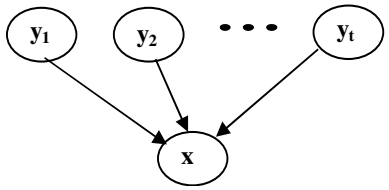
$$P(H|e) = \frac{P(e|H)\,P(H)}{P(e)}$$

where, H : a hypothesis
e : is an evidence
P(H) : prior probability
P(H|e) : posterior probability
P(e|H) : probability of e if H is true
P(e) : a normalizing constant, then
we write: P(H|e) ~ P(e|H)P(H)

Bayesian Networks

Bayesian networks are **directed acyclic graphs (DAGS)** in which the **nodes** represent *random variables*, the **arcs** portray *causal relationships* between these variables, and the **strengths** of these causal influences are expressed by *conditional probabilities*.

y_i : parent nodes (in this case, root nodes)
x : child node
y_i cause x
Y the set of parents of x
The influence of Y on x
can be quantified by any function
F(x,Y) such that $\Sigma_{\forall x} F(x,Y) = 1$
$0 \leq F(x,Y) \leq 1$
For example, F(x,Y)=P(x|Y)



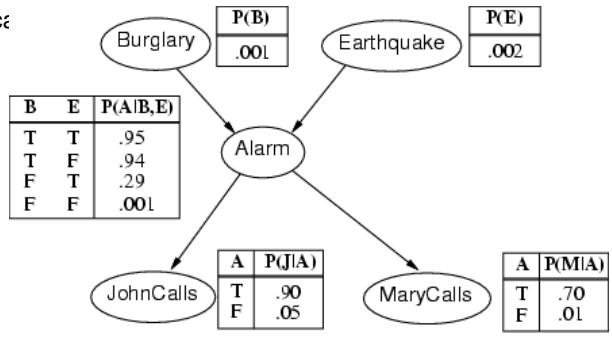
Παράδειγμα

I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

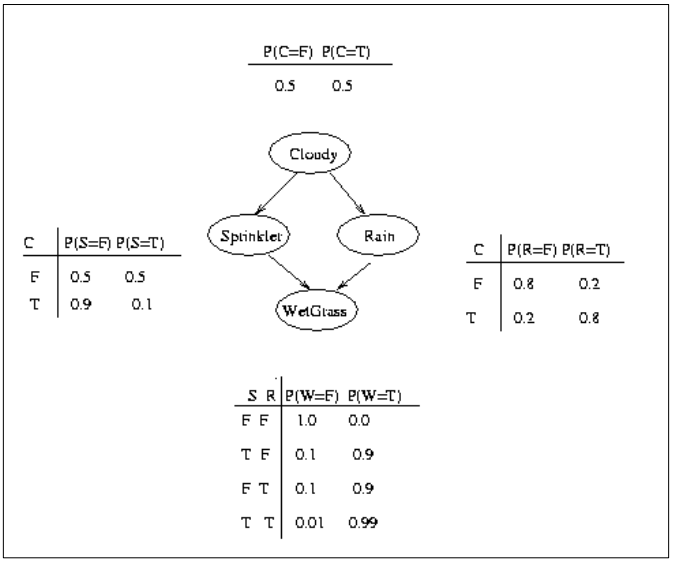
Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

Network topology reflects "causal" knowledge:

- A burglar can cause the alarm
- An earthquake can set the alarm
- The alarm can cause Mary to call
- The alarm can cause John to ca



Παράδειγμα

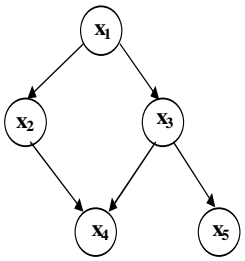


Bayesian Networks and Joint Probabilities

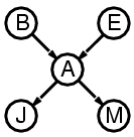
The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(x_1, \ldots, x_n) = \Pi_{i=1,n}\, \mathbf{P}(x_i \mid Parents(x_i))$$

Examples:



$\mathbf{P}(x_1, x_2, x_3, x_4, x_5) =$
$\mathbf{P}(x_1)\, \mathbf{P}(x_2 \mid x_1)\, \mathbf{P}(x_3 \mid x_1)\, \mathbf{P}(x_4 \mid x_2, x_3)\, \mathbf{P}(x_5 \mid x_3)$.
$P(x_1)$: prior probability of the root node



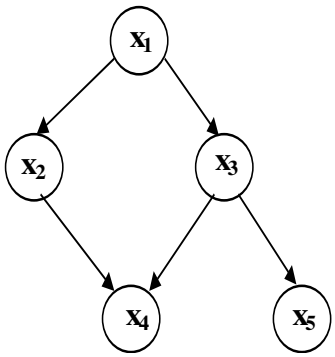
$\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e) =$
$\mathbf{P}(j \mid a)\, \mathbf{P}(m \mid a)\, \mathbf{P}(a \mid \neg b, \neg e)\, \mathbf{P}(\neg b)\, \mathbf{P}(\neg e)$

Also

In a Bayesian network each variable x is conditionally independent of all its non-descendants, given its parents. (ανεξάρτητης των μη απογόνων δοθέντων των πατέρων)

For example:

$\mathbf{P}(x_4, x_5 \mid x_2, x_3) = \mathbf{P}(x_4 \mid x_2, x_3)\, \mathbf{P}(x_5 \mid x_3)$



Information Retrieval Models

Inference Network Model

Inference Network Model

- Επιστημολογική άποψη του προβλήματος της ΑΠ
- Σχολές σκέψης στις πιθανότητες:
  - *freqüentist:*

    πιθανότητα = στατιστικό μέγεθος σχετιζόμενο με την τύχη
  - *epistemological*

    πιθανότητα = βαθμός πίστης του οποίου η περιγραφή μπορεί να είναι απαλλαγμένη από στατιστικά πειράματα
- Η προσέγγιση του Inference Network Model:
  - Επιστημολογική χρήση των πιθανοτήτων

Inference Network Model: Τυχαίες μεταβλητές

- Εγγράφων
  - Η τυχαία μεταβλητή που σχετίζεται με ένα έγγραφο d_j παριστάνει το **συμβάν παρατήρησης** του εγγράφου αυτού
- Όρων
  - Η παρατήρηση ενός εγγράφου είναι η αιτία για **αυξημένη πίστη** στις τυχαίες μεταβλητές που αντιστοιχούν στους όρους που περιέχει το κείμενο
- Επερωτήσεων
  - εκφράζει το βαθμό ικανοποίησης της επερώτησης

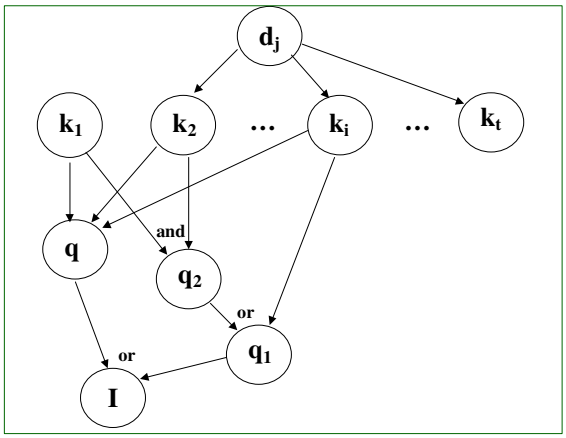
- Όλες οι τυχαίες μεταβλητές (d,k,q) είναι **δυαδικές** (0 ή 1) **!**

Inference Network Model: Δομή
(documents => terms => queries)

d_j has index terms k_2, k_i, and k_t
q has index terms k_1, k_2, and k_i
q_1 and q_2 model boolean formulation
$q_1=((k_1 \Lambda k_2)$ v $k_i)$;
I = (q v q_1)

Nodes

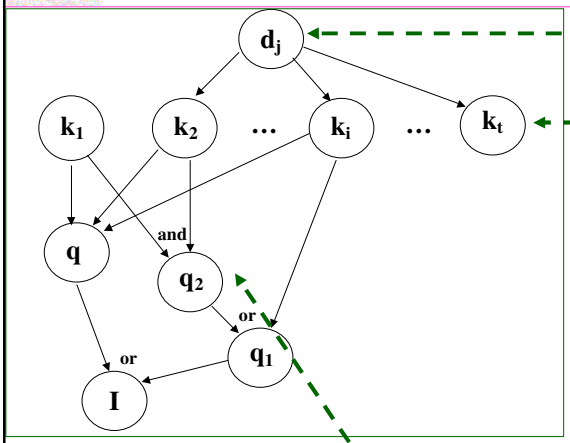
documents (d_j)
index terms (k_i)
queries (q, q_1, and q_2)
user information need (I)



Inference Network Model: Ranking

- Όλες οι τυχαίες μεταβλητές (d,k,q) είναι **δυαδικές** (0 ή 1) **!**
- **Κατάταξη** Rank(q,dj)
  - Rank(q,dj) = P(q Λ d_j)
  - q and dj είναι συντομογραφίες του q=1 and dj =1
  - Rank(q,dj) :
    - expresses how much evidential support the observation of dj provides to the query q
  - Rank(q,dj) = P(q=1 Λ d_j=1)
  - (dj stands for a state where dj = 1 and $\forall i \neq j \Rightarrow$ di =0, because we observe one document at a time)

How an inference network can be tuned to subsume the Boolean Model



P(dj) = 1/N

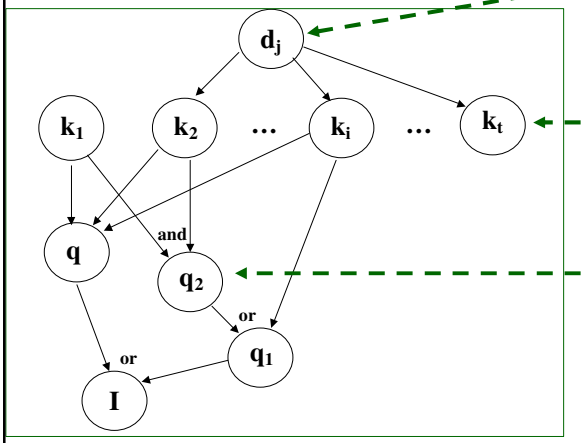
P(k_i | d_j) = 1 if wij=1

P(q | **k**) =1 if $\exists q_{cc}$ | ($q_{cc} \in q_{dnf}$) Λ ($\forall$ k_i, $g_i(k) = g_i(q_{cc})$

0 otherwise

$\Rightarrow$ one of the conjunctive components of the query **must be** matched by the active index terms in **k**

Inference network and TF-IDF



$P(dj) = 1/|\mathbf{d_j}|$

$\Rightarrow$ reflects the importance of document **normalization**

$P(k_i \mid d_j) = tf_{i,j}$

$P(q \mid ki) = idf_i$

$\Rightarrow$ however the obtained ranking is distinct of the one provided by the vector model

Inference Network Model: Σύνοψη

- Επιστημολογική προσέγγιση
- Τοπολογία δικτύου d -> k -> q
- Εκφραστική ικανότητα
  - Συλλαμβάνει το Boolean Model (με δυαδικές τυχαίες μεταβλητές)
  - Μπορεί να προσεγγίσει αρκετά το διανυσματικό
  - Δυνατότητα συνδυασμού πολλών αποδεικτικών πηγών
- Συστήματα βασισμένα σε αυτό το μοντέλο
  - Inquery system

Information Retrieval Models

Belief Network Model

Belief Network Model

- Όπως και στο Inference Network Model
  - Επιστημολογική άποψη του προβλήματος της ΑΠ
  - Τυχαίες μεταβλητές για έγγραφα, όρους και επερωτήσεις
- Αντίθετα με το Inference Network Model, εδώ έχουμε
  - πιο ξεκάθαρο δειγματικό χώρο
  - συνολοθεωρητική προσέγγιση
  - διαφορετική τοπολογία δικτύου

Belief Network Model: The Probability Space

Δειγματικός Χώρος: **K**={k_1, k_2, ...,k_t} (sample space)

Κάθε όρος ki είναι μια στοιχειώδης έννοια (elementary concept)
Κάθε σύνολο u $\subset$ **K** είναι μια έννοια

Σε κάθε u $\subset$ **K** αντιστοιχίζεται ένα δυαδικό διάνυσμα **k**=(k_1, k_2, ...,k_t)
τ.ω. w_i=1 $\Leftrightarrow$ $k_i \in u$

k_i a binary random variable associated with the index term k_i, (k_i = 1 $\Leftrightarrow$ w_i=1 $\Leftrightarrow$ $k_i \in u$)

Belief Network Model: Documents and Queries

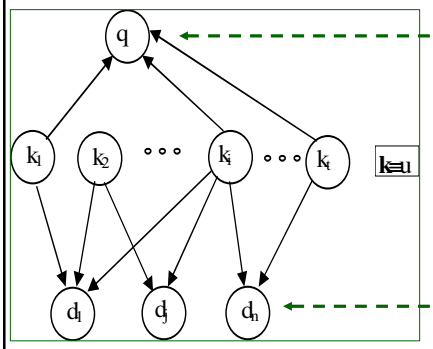
Εγγραφο d:
ένα υποσύνολο του K (που περιέχει τους όρους του d)

Επερώτηση q:
ένα υποσύνολο του K (που περιέχει τους όρους της q)

Ορίζουμε μια κατανομή πιθανότητας P στο K ως εξής:
$P(c)=\Sigma_{\forall u} P(c|u)\ P(u)$ // the *degree of coverage* of the space K by c
c: a generic concept representing a document or a query

Αρχικά (που δεν γνωρίζουμε τα P(u)), υποθέτουμε ότι:
$P(u)=(1/2)^t = 1/2^{|K|}$

Belief Network Model: Τοπολογία Δικτύου και Σημασιολογία



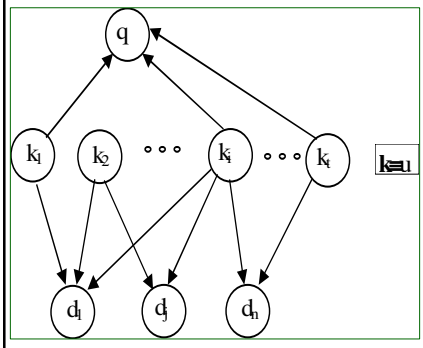
P(q)=1 means that q covers completely the concept space

P(dj)=1 means that dj covers completely the concept space

So user queries and documents are modeled as **subsets** of index terms.

The concept space K works as the common sample space.

Belief Network Model: **Ranking : P(d|q)**

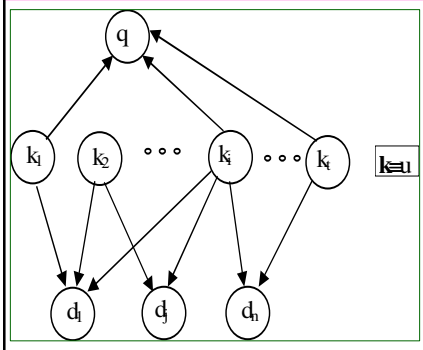


P(d | q)

Εκφράζει το βαθμό κάλυψης του d από την την επερώτηση q

Όσο πιο μεγάλος είναι, τόσο πιο συναφές είναι το d με το q

Belief Network Model: **Computing** P(d|q)



P(d | q)

Εκφράζει το βαθμό κάλυψης του d από την την επερώτηση q

$P(d_j|q) = P(d_j \Lambda q) / P(q)$ (εξ΄ορισμού: θεώρημα Bayes)

$\sim P(d_j \Lambda q)$ (διότι ο παρανομαστής P(q) είναι ίδιος για όλα τα d)

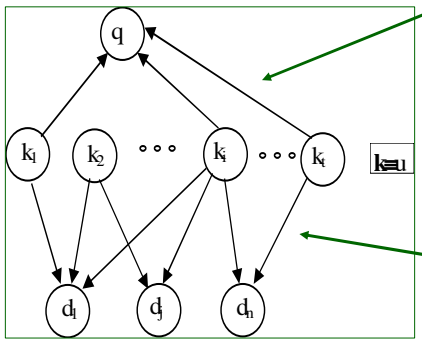
$\sim \Sigma_{\forall u} P(d_j \Lambda q \mid u) P(u)$ (εξ΄ορισμού)

$\sim \Sigma_{\forall u} P(d_j \mid u)\ P(q \mid u) P(u)$ (Bayesian networks)

$\sim \Sigma_{\forall \mathbf{k}} P(d_j \mid \mathbf{k})\ P(q \mid \mathbf{k}) P(\mathbf{k})$

Αρκεί να ορίσουμε αυτά

Defining P(q|k) and P(d|k) to model the vector model



$$P(q \mid k_i) = \frac{w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{iq}^2}} \text{ if } q \text{ contains } k_i \text{ (otherwise} = 0)$$

$$P(d_j \mid k_i) = \frac{w_{ij}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}} \text{ if } d_j \text{ contains } k_i \text{ (otherwise} = 0)$$

In this way we get the ranking of the classic Vector Space Model !

Belief Network Model: Σύνοψη

- Επιστημολογική προσέγγιση
- Τοπολογία δικτύου d <- k -> q
  - (το inference network είχε δομή d -> k -> q)
- Εκφραστική ικανότητα
  - Συλλαμβάνει το Boolean Model (με δυαδικές τυχαίες μεταβλητές)
  - Συλλαμβάνει το Διανυσματικό Μοντέλο
  - Μπορεί αναπαράγει οποιαδήποτε διάταξη του Inference Network (το αντίστροφο δεν ισχύει)
  - Δυνατότητα συνδυασμού πολλών αποδεικτικών πηγών
- Inference Network Model:
  - το πρώτο και πιο γνωστό
  - επιτυχημένη χρήση στο σύστημα Inquery

Διάρθρωση

PART (A)

- *Ανάκτηση και Φιλτράρισμα*
- *Εισαγωγή στα Μοντέλα Άντλησης*
- *Κατηγορίες Μοντέλων*
- *Exact vs Best Match*
- *Τα κλασσικά μοντέλα ανάκτησης*
  - *Το Boolean Μοντέλο*
  - *Στατιστικά Μοντέλα - Βάρυνση Όρων*
  - *Το Διανυσματικό Μοντέλο*
  - *Το Πιθανοκρατικό Μοντέλο*

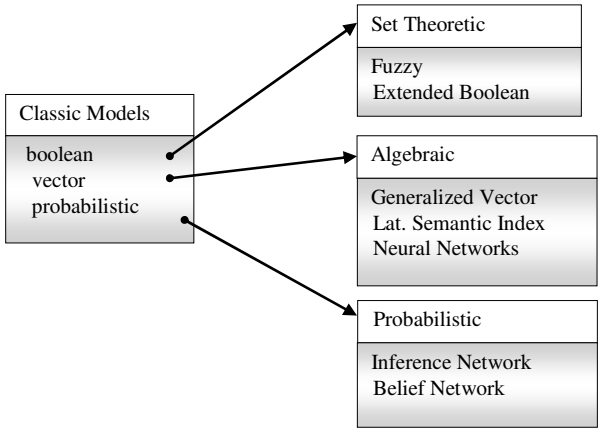
PART (B):***Εναλλακτικά μοντέλα***

- *(I) Συνολοθεωρητικά μοντέλα*
  - *Fuzzy Retrieval Model*
  - *Extended Boolean Model*
- *(II) Αλγεβρικά Μοντέλα*
  - *Latent Semantic Indexing*
  - *Neural Netwok Model*

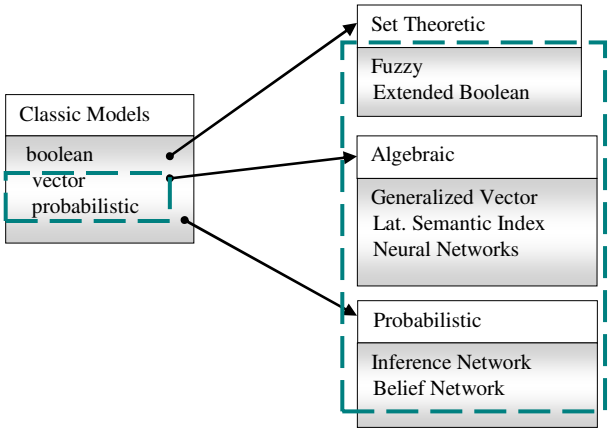
PART (C):

- *(III) Πιθανοκρατικά Μοντέλα*
  - *Bayesian Network Model*
  - *Inference Network Model*

Ταξινομία Μοντέλων που εξετάσαμε

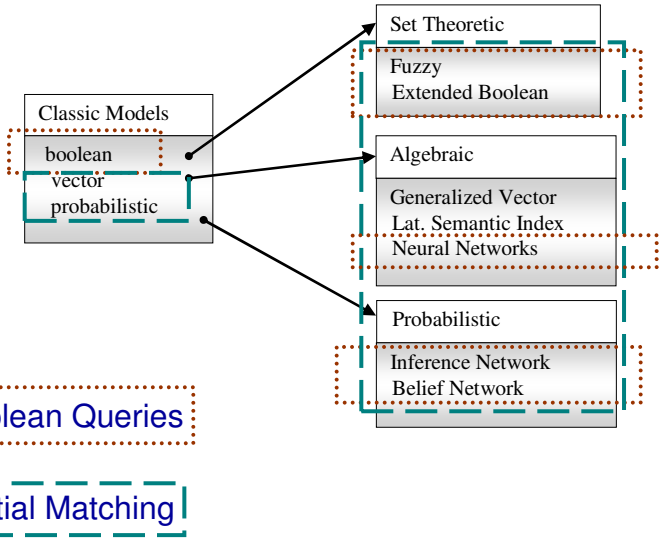


Ταξινομία Μοντέλων που εξετάσαμε

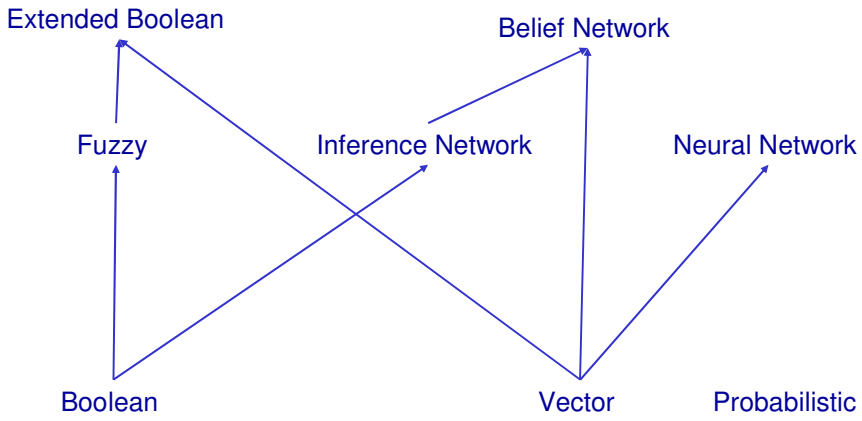


Partial Matching

Ταξινομία Μοντέλων που εξετάσαμε



WRT Expressive Power (incomplete)



Άλλοι τύποι Μοντέλων Ανάκτησης που θα δούμε αργότερα

Αργότερα

- Μοντέλα Ανάκτησης **Δομημένων** Εγγράφων
  - Non Overlapping Lists
  - Proximal Nodes
  - Retrieval Models for XML
- Μοντέλα Ανάκτησης **Ιστοσελίδων**
  - Έμφαση στους συνδέσμους
- Μοντέλα Ανάκτησης **Πολυμέσων**
- Μοντέλα Βασισμένα στη **Λογική**
  - Carlo Meghini and Umberto Straccia, A Relevance Terminological Logic for Information Retrieval,Proceedings of SIGIR'96, Zurich, Switzerland, 1996