Recuperação de Informação B

Cap. 02: Modeling (Extended Boolean Model)

2.6.2

September 13, 1999
Extended Boolean Model

- Boolean model is simple and elegant.
- But, no provision for a ranking
- As with the fuzzy model, a ranking can be obtained by relaxing the condition on set membership
- Extend the Boolean model with the notions of partial matching and term weighting
- Combine characteristics of the Vector model with properties of Boolean algebra
The Idea

- The *extended Boolean model* (introduced by Salton, Fox, and Wu, 1983) is based on a critique of a basic assumption in Boolean algebra.

- Let,
  - $q = kx \land ky$
  - $wx_j = f_{xj} \times \frac{idf(x)}{\max(idf(i))}$ associated with $[kx,dj]$
  - Further, $wx_j = x$ and $wy_j = y$
The Idea:

- $q_{\text{and}} = k_x \land k_y$; $w_{xj} = x$ and $w_{yj} = y$

Sim($q_{\text{and}}, d_j$) = $1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$
The Idea: $qor = kx \lor ky; \quad wxj = x \text{ and } wyj = y$

$$\text{sim}(qor,dj) = \sqrt{\left(\frac{x^2 + y^2}{2}\right)}$$
Generalizing the Idea

- We can extend the previous model to consider Euclidean distances in a t-dimensional space.
- This can be done using *p-norms* which extend the notion of distance to include $p$-distances, where $1 \leq p \leq \infty$ is a new parameter.
- A generalized conjunctive query is given by
  $\land p^ 1 \lor p^ 2 \lor p \ldots \lor p^ t$
- A generalized disjunctive query is given by
  $\lor p^ 1 \land p^ 2 \land p \ldots \land p^ t$
Generalizing the Idea

- $\text{sim}(qor,dj) = \left( x_1^p + x_2^p + \ldots + x_m^p \right)^{\frac{1}{p}}$

- $\text{sim}(q\text{and},dj) = 1 - \left( (1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p \right)^{\frac{1}{p}}$
Properties

- $\text{sim}(q_{or},dj) = \left(\frac{x_1^p + x_2^p + \ldots + x_m^p}{m}\right)^{\frac{1}{p}}$

- If $p = 1$ then (Vector like)
  - $\text{sim}(q_{or},dj) = \text{sim}(q_{and},dj) = \frac{x_1 + \ldots + x_m}{m}$

- If $p = \infty$ then (Fuzzy like)
  - $\text{sim}(q_{or},dj) = \max (wx_j)$
  - $\text{sim}(q_{and},dj) = \min (wx_j)$
Properties

- By varying $p$, we can make the model behave as a vector, as a fuzzy, or as an intermediary model.
- This is quite powerful and is a good argument in favor of the extended Boolean model.
- $(k_1 \lor^2 k_2) \land^\infty k_3$

- $k_1$ and $k_2$ are to be used as in a vectorial retrieval while the presence of $k_3$ is required.
Properties

- \( q = (k_1 \vee^2 k_2)^\wedge\infty k_3 \)
- \(\text{sim}(q,dj) = (\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2} \right)^\frac{1}{p}\right) + x_3^p)^\frac{1}{p}\)
Conclusions

- Model is quite powerful
- Properties are interesting and might be useful
- Computation is somewhat complex
- However, distributivity operation does not hold for ranking computation:
  - $q_1 = (k_1 \lor k_2) \land k_3$
  - $q_2 = (k_1 \land k_3) \lor (k_2 \land k_3)$
  - $\text{sim}(q_1,dj) \neq \text{sim}(q_2,dj)$