정보검색 모델
Extended Boolean Model

국민대학교 컴퓨터공학부

강 승 식

IR Models

- Set theoretic models: Boolean
  - Fuzzy set model
  - Extended Boolean model
- Algebraic models: Vector
  - Vector space model
  - Latent Semantic Index
  - Neural Network
- Probabilistic models: Probabilistic
  - Inference Network
  - Belief Network

Retrieval

- Ad hoc
  - The documents in the collection remain relatively static,
  - while new queries are submitted to the system

- Filtering
  - The queries remain relatively static,
  - while new documents come into the system (and leave)
    - Ex) 매일 아침 신문기사 중에서 내가 관심있는 분야의 keyword에 대한 검색: 주식, 축구, 야구 등

Boolean Model

- Definition

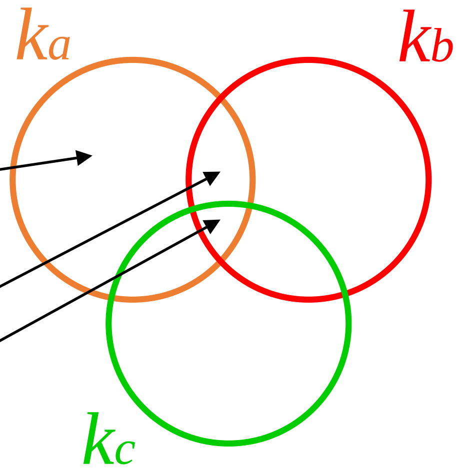
$$\vec{d}_j = (w_{1j}, w_{2j}, \ldots w_{tj}) \qquad w_{ij} \in \{0,1\}$$

$$sim(d_j, q) = \begin{cases} 1, & if\ \exists \vec{q}_{cc} \mid (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall k_i, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0, & otherwise \end{cases}$$

- Example

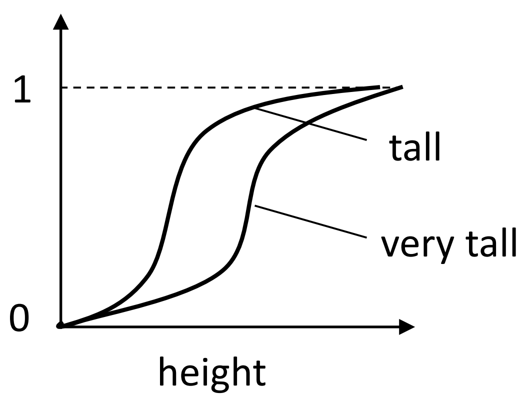
$$q = k_a \wedge (k_b \vee \neg k_c)$$

$$\vec{q}_{dnf} = (1,1,1) \vee (1,1,0) \vee (1,0,0)$$

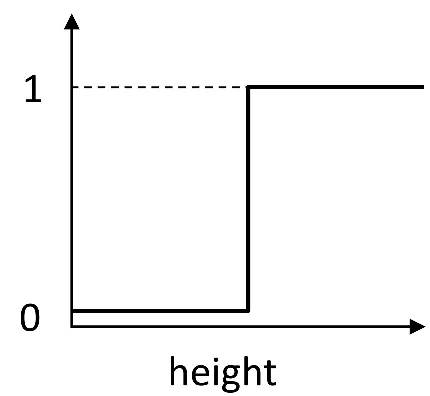


Fuzzy Set Model

- Fuzzy Set Theory
  - Deals with the representation of classes whose boundaries are not well defined
  - <u>Membership</u> in a fuzzy set is a notion intrinsically <u>gradual</u> instead of abrupt (as in conventional Boolean logic)



Fuzzy Membership



Conventional Membership

Fuzzy Set: definition

- Fuzzy set *A* of a universe of discourse *U*
- Membership function

$$\mu_A : U \rightarrow [0,1]$$

- For each element *u* of *U*,

$$\mu_A(u)$$

- Membership value of *u* is calculated by

$$\mu_{\overline{A}}(u) = 1 - \mu_A(u)$$

$$\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$$

$$\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$$

Fuzzy Set Model

- Fuzzy information retrieval
  - Representing documents and queries through sets of keywords yields descriptions which are only <u>partially related</u> to the real semantic contents of the respective documents and queries
  - Each *query term* defines a *fuzzy* set
  - Each *document* has a <u>*degree of membership*</u> in this set
- Rank the documents relative to the user query

$$\mathrm{D_t} = \{(d_1, 0.8), (d_2, 0.5)\},\ \ \mathrm{D_s} = \{(d_1, 0.5), (d_2, 0.4)\}$$

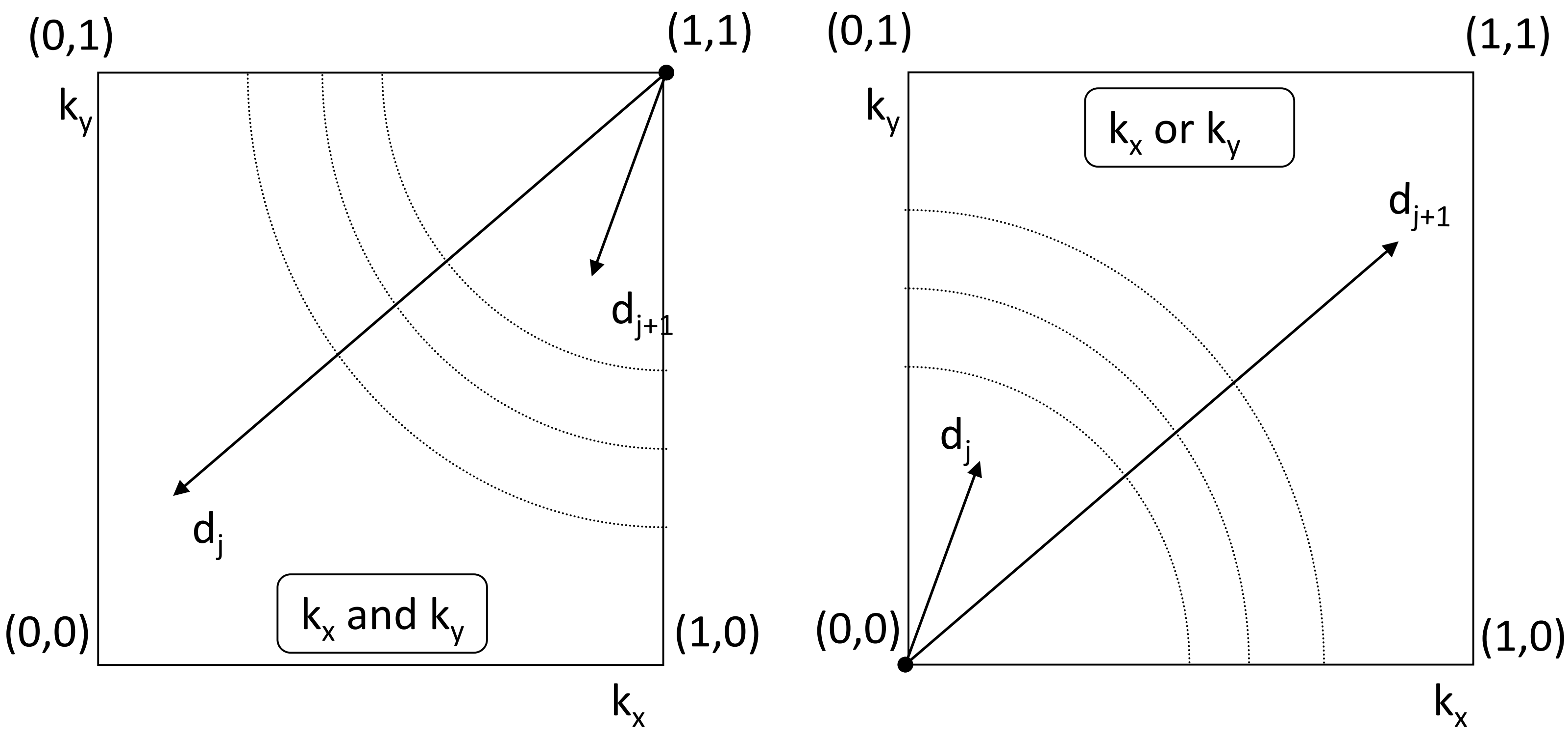
$$\mathrm{Q(s \vee t) = D_s \cup D_t = \{(d_1, 0.8), (d_2, 0.5)\}}$$
$$\mathrm{Q(s \wedge t) = D_s \cap D_t = \{(d_1, 0.5), (d_2, 0.4)\}}$$

Extended Boolean Model

- Critique of a basic assumption of Boolean logic
  - Conjunction Boolean query : $q = k_x \wedge k_y$
    - 1개만 포함된 문서 vs. 둘 다 포함 안된 문서 → 모두 동일하게 취급
  - Disjunction Boolean query : $q = k_x \vee k_y$
    - 1개만 포함된 문서 vs. 둘 다 포함된 문서 → 모두 동일하게 취급

- *Combine* Boolean query formulations with characteristics for the vector model

- When only two terms are considered, queries and documents are plotted in a two dimensional map



- Disjunctive query : $q_{or} = k_x \vee k_y$
  - Point (0,0) is the spot to be avoided
  - Measure of similarity
    - Distance from the point (0,0)

$$sim(q_{or}, d) = \sqrt{\frac{x^2 + y^2}{2}}$$

- Conjunctive query : $q_{and} = k_x \wedge k_y$
  - Point (1,1) is the most desirable spot
  - Measure of similarity
    - Complement of the distance from the point (1,1)

$$sim(q_{and}, d) = 1 - \sqrt{\frac{(1-x)^2 + (1-y)^2}{2}}$$

P-norm Model

- *Generalizes* the notion of distance to include not only Euclidean distance but also *p*-distances
  - *p* value is specified at query time
- Generalized disjunctive query

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$$

- Generalized conjunctive query

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

Similarity Measure

- Similarity measure

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots \vee^p k_m$$

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m$$

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

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

- Example

$$q = (k_1 \wedge^p k_2) \vee^p k_3$$

$$sim(q, d_j) = \left(\frac{\left(1 - \left(\frac{(1-x_1)^p + (1-x_2)^p}{2} \right)^{\frac{1}{p}} \right)^p + x_3^p}{2} \right)^{\frac{1}{p}}$$

Ranking functions

- Waller-Kraft, Paice, P-Norm and Infinite –One

$$F(d, t_1 \text{ AND } t_2) = (1-\gamma) \bullet \text{MIN}(w_{d1}, w_{d2}) + \gamma \bullet \text{MAX}(w_{d1}, w_{d2}), \quad 0 \le \gamma \le 0.5$$

$$F(d, t_1 \text{ OR } t_2) = (1-\gamma) \bullet \text{MIN}(w_{d1}, w_{d2}) + \gamma \bullet \text{MAX}(w_{d1}, w_{d2}), \quad 0.5 \le \gamma \le 1$$

(a) The Waller-Kraft model

$$F(d, t_1 \text{ AND } t_2) = \frac{1}{1+r} \bullet \text{MIN}(w_{d1}, w_{d2}) + \frac{r}{1+r} \bullet \text{MAX}(w_{d1}, w_{d2}), \quad 0 \le r \le 1$$

$$F(d, t_1 \text{ OR } t_2) = \frac{1}{1+r} \bullet \text{MAX}(w_{d1}, w_{d2}) + \frac{r}{1+r} \bullet \text{MIN}(w_{d1}, w_{d2}), \quad 0 \le r \le 1$$

(b) The Paice model

$$F(d, t_1 \text{ AND } t_2) = 1 - \left[\frac{(1-w_{d1})^p + (1-w_{d2})^p}{2}\right]^{1/p}, \quad 1 \le p \le \infty$$

$$F(d, t_1 \text{ OR } t_2) = \left[\frac{w_{d1}{}^p + w_{d2}{}^p}{2}\right]^{1/p}, \quad 1 \le p \le \infty$$

(c) The P-Norm model

$$F(d, t_1 \text{ AND } t_2) = \gamma \bullet (1 - \text{MAX}(1-w_{d1}, 1-w_{d2})) + (1-\gamma) \bullet \frac{w_{d1} + w_{d2}}{2}, \quad 0 \le \gamma \le 1$$

$$F(d, t_1 \text{ OR } t_2) = \gamma \bullet \text{MAX}(w_{d1}, w_{d2}) + (1-\gamma) \bullet \frac{w_{d1} + w_{d2}}{2}, \quad 0 \le \gamma \le 1$$

(d) The Infinite-One model

Figure 2. Binary evaluation formulas for AND and OR operations

Weighting Scheme

- Term Frequency (*tf*)
  - Measure of *how well that term describes the document* contents

$$f_{ij} = \frac{freq_{ij}}{\max_l \ freq_{lj}} \quad (freq_{ij} : \text{Raw frequency of term } k_i \text{ in the document } d_j)$$

- Inverse Document Frequency (*idf*)
  - *Terms which appear in many documents are not very useful for distinguishing* a relevant document from a non-relevant one

$$idf_i = \log \frac{N}{n_i}$$

n_i : Number of documents in which the index term k_i appears

N : Total number of documents

Vector Space Model

- Definition

$$\vec{q} = (w_{1q}, w_{2q}, \ldots, w_{tq}) \qquad w_{iq} \geq 0$$

$$(w_{ij} = tf \times idf)$$

$$\vec{d}_j = (w_{1j}, w_{2j}, \ldots, w_{tj}) \qquad w_{ij} \geq 0$$

$$sim(d_j, q) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{ij} \times w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^{2}} \times \sqrt{\sum_{i=1}^{t} w_{iq}^{2}}}$$

$$0 \leq sim(d_j, q) \leq 1 \qquad \text{(cosine similarity)}$$

$|\vec{q}|$: Does not affect the ranking

$|\vec{d}_j|$: Normalization in the space of the documents

References

- Joon-Ho Lee, “Properties of Extended Boolean Models in Information Retrieval,” SIGIR-94, pp.182-190, 1994.
- W. Waller and D. Kraft, “A Mathematica Model of a Weighted Boolean Retrieval System,” Information Processing & Management, pp.235-245, 1979.
- C. Paice, “Soft Evaluation of Boolean Search Queries in Information Retrieval Systems,”, Information Technology: Research and Development, Vol.3, No.1, pp.33-42, 1984.