Fuzzy Set Model

Fuzzy Set:

Let $X$ be the universe of discourse, with elements of $X$ denoted as $x$. A fuzzy set $A$ of $X$ is characterized by a membership function $\mu_A(x)$ that associates each element $x$ with a degree of membership value in $A. 0 \leq \mu_A(x) \leq 1.$
**Fuzzy Set Example**

*Young* is fuzzy set. Its membership function could be defined as the follows:

![Membership Function Diagram](image)
Fuzzy Set Operations

• Intersection
  \[ \mu_{A \cap B}(x) = \min\{\mu_A(x), \mu_B(x)\} \]
  \[ \mu_{A \cap B}(x) = \mu_A(x) \times \mu_B(x) \]

• Union
  \[ \mu_{A \cup B}(x) = \max\{\mu_A(x), \mu_B(x)\} \]
  \[ \mu_{A \cup B}(x) = 1 - (1 - \mu_A(x)) \times (1 - \mu_B(x)) \]

• Complement
  \[ \mu_{\neg A}(x) = 1 - \mu_A(x) \]
Term Correlation Matrix: C

- C is a t by t matrix, and each column (row) corresponds to an index term.

- $c_{ij} = \frac{n_{ij}}{n_i + n_j - n_{ij}}$
  - $n_i$ is the number of documents containing the term $k_i$ and $n_{ij}$ is the number of documents containing both $k_i$ and $k_j$.

- A larger value of $c_{ij}$ means the two terms are more correlated.
Fuzzy Sets of Documents Defined by Index Terms

Using the correlation matrix, each index term $k_i$ defines a fuzzy set of documents. The membership function for a document $d_j$ is:

$$
\mu_{ij} = 1 - \prod_{k_l \in d_j} (1 - c_{il})
$$
Document And Query Representation

- Document
  a list of index terms contained in the document

- Query
  a logic expression same as in the boolean model
Similarity Measure

Relevant documents of a query is a fuzzy set defined by its membership function.

Example:

Query: $k_i$

$$\mu_{ij} = 1 - \prod_{k_l \in d_j} (1 - c_{il})$$
More Query Examples

- $q = k_1$ and $k_2$
  \[
  \mu_{qj} = \mu_{1j} \times \mu_{2j} = (1 - \prod_{k_1 \in d_j} (1 - c_{1_{k_1}})) \times (1 - \prod_{k_2 \in d_j} (1 - c_{2_{k_2}}))
  \]

- $q = k_1$ or $k_2$
  \[
  \mu_{qj} = 1 - (1 - \mu_{1j}) \times (1 - \mu_{2j}) = 1 - (1 - (1 - \prod_{k_1 \in d_j} (1 - c_{1_{k_1}}))) \times (1 - (1 - \prod_{k_2 \in d_j} (1 - c_{2_{k_2}}))))
  \]
  \[
  = 1 - \prod_{k_1 \in d_j} (1 - c_{1_{k_1}}) \times \prod_{k_2 \in d_j} (1 - c_{2_{k_2}})
  \]
Advantages And Disadvantages

• Advantages
  – use fuzzy set theory. The set of relevant documents is fuzzy.
  – query terms which are in a document have aslo impact.

• Disadvantages
  – no term frequency
  – not well evaluated
Extended Boolean Model

• Boolean Model:
  – query: logic expression
  – document: binary representation
  – similarity: simple binary similarity

• Extended Boolean Model:
  – query: logic expression
  – document: vector representation with term weighting
  – match: generalized similarity measure
Generalized Similarity Measure

• *And* logic Operation
  relevant documents should be close to (1 1 . 1)

• *Or* logic Operation
  relevant documents should be far from (0 0 . 0)

• Different similarity measures for *And* and *Or* operations.
Generalized Similarity Measure

Assume that all term weights are in $[0, 1]$.

- $q = k_1$ and $k_2$ ($w_1$ and $w_2$ are their weights in document $d$)

$$sim(q,d) = 1 - \sqrt{\frac{w_1^2 + w_2^2}{2}}$$

- $q = k_1$ or $k_2$

$$sim(q,d) = 1 - \sqrt{\frac{(1-w_1)^2 + (1-w_2)^2}{2}}$$
P-norm Model

Assume $x_i$ is the term weights for terms in the query and $m$ is the number of terms in the query

$$sim(q_{or}, d) = \left( \frac{1}{m} \sum_{i=1}^{m} x_i^p \right)^{\frac{1}{p}}$$

$$sim(q_{and}, d) = 1 - \left( \frac{1}{m} \sum_{i=1}^{m} (1-x_i)^p \right)^{\frac{1}{p}}$$

$$sim(q_{not}, d) = 1 - sim(q, d)$$
P-Norm

- $P$ is the degree of strictness
- $P = 1$: normal vector model and least strictness
- $P = \infty$: boolean model and most strictness
- $P = 2$: extended boolean model
Characteristics

• Boolean logic queries

• Ranked retrieval

• $tf*idf$ weighting

• Effective
Latent Semantic Indexing

• User singular value decomposition (a dimensionality reduction technique) to identify uncorrelated, significant basis vector or factor

• Replace original index terms with a subset of new factors (concepts) in both documents and queries

• Compute the similarity in this new space