Jul – Nov 2022
IT458

Classical IR Models for Unstructured Text

Fuzzy Set Models

Beyond the Boolean Model

- Characteristics of the Boolean model –
  - uses binary criterion for deciding relevance
  - No support for partial matching
  - No ranking.
- Led to development of newer Set theory-based models
  - A popular set theoretic model: Fuzzy Set Model

Fuzzy Logic - Basics

What is Fuzzy Logic?

- A type of logic that recognizes more than simple true and false values.
  - propositions can be represented with **degrees** of **truthfulness** and **falsehood**.

- **Example: "Today is sunny"**
  - 100% true if there are no clouds
  - 0% true if it rains all day

What is Fuzzy Logic?

- A type of logic that recognizes more than simple true and false values.
  - propositions can be represented with **degrees** of **truthfulness** and **falsehood**.

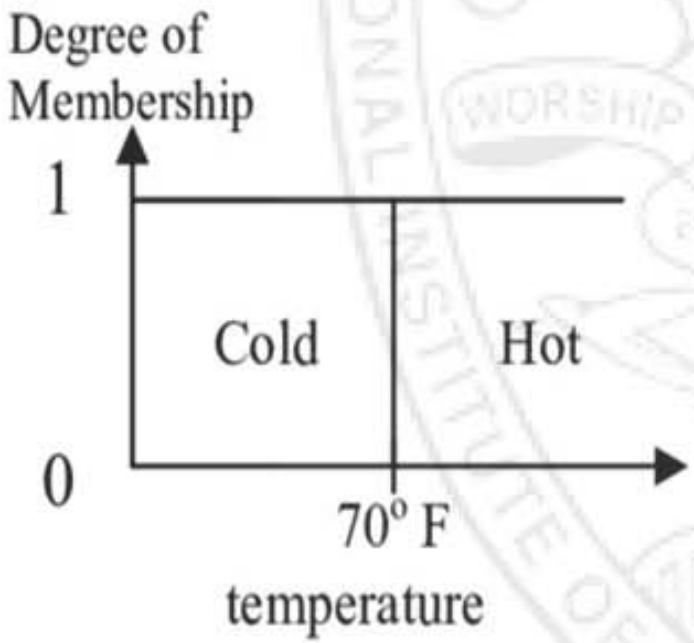
- **Example: "Today is sunny"**
  - **100% true if there are no clouds**
  - 80% true if there are a few clouds
  - 50% true if it's hazy
  - 20% true if its drizzles sporadically.
  - **0% true if it rains all day**

Fuzzy Logic - Basics

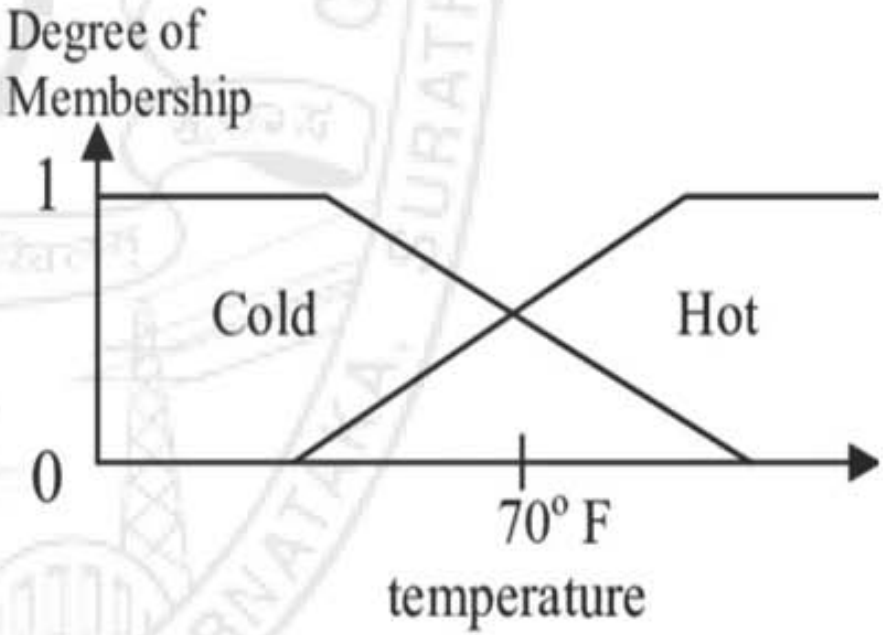
- Fuzzy Set (or *Uncertain sets*)
  - Similar to classical sets, but elements have degrees of membership.
- Membership Function
  - A function defined on a fuzzy set (with a range in the interval [0,1]), covering all possible values in the given domain.
- Degree of Membership
  - indicates the level of belongingness of a particular item to a fuzzy set.
  - A value in the range [0,1]

Fuzzy Logic - Basics

Crisp set



Fuzzy set



Fuzzy Set Model for IR

Fuzzy Set Model for IR

- Based on **Fuzzy Sets** (Zadeh, 1965)
  - a mathematical system that captures aspects of the ambiguity of human language and thought.
  - solved problems in areas such as artificial intelligence and the automated control of machines.

Fuzzy Set Model for IR

- *Key idea:* notion of a degree of membership associated with the elements of a set *(i.e. terms in a document / query)*
  - Define a framework for representing classes whose boundaries are not well defined.
  - degree of membership varies from 0 to 1 and allows modelling the notion of **marginal membership**

Formalisms – Fuzzy Set IR Model

- Document d_i - CNF of weighted index terms

D_1 = (AND <apple, 0.4> <orange, 0.3> <banana, 0.5>);

Formalisms – Fuzzy Set IR Model

- Document d_i - CNF of weighted index terms

  D_1 = (AND <apple, 0.4> <orange, 0.3> <banana, 0.5>);

- Query q_j - as in the Boolean Model (using Boolean operators)

  Q = (apple OR orange)

- Matching function – *Retrieval Status Value (RSV)*
  - computes the degree to which document d_i satisfies q_j

Fuzzy Set Model for IR

- *Definition*:

  Let **U** be the universe of discourse, **A** and **B** be two fuzzy subsets of U. Let **u** be an element of U. Then,

$$\mu_{\overline{A}} = 1 - \mu_A(u) \quad \text{[COMPLEMENT]}$$

$$\mu_{A \cup B}(u) = \max\{ \mu_A(u), \mu_B(u) \} \quad \text{[UNION]}$$

$$\mu_{A \cap B}(u) = \min\{ \mu_A(u), \mu_B(u) \} \quad \text{[INTERSECTION]}$$

Fuzzy Set Model for IR - Example

Example :

- Let $V = \{k_1, k_2, k_3, k_4, k_5, k_6\}$

  $A = \{k_1, k_2, k_3\}$ $B = \{k_2, k_3, k_4\}$

- Assume $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

  $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

$$\mu_{A \cup B}(u) =$$

$$\mu_{A \cap B}(u) =$$

$$\mu_{\bar{A}} =$$

$$\mu_{\bar{B}} =$$

Fuzzy Set Model for IR - Example

Example :

- Let $V = \{k_1, k_2, k_3, k_4, k_5, k_6\}$

  $A = \{k_1, k_2, k_3\}$ $B = \{k_2, k_3, k_4\}$

- Assume $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

  $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

$$\mu_{A \cup B}(u) = \{k_1:0.8, k_2:0.7, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

$$\mu_{A \cap B}(u) =$$

$$\mu_{\bar{A}} =$$

$$\mu_{\bar{B}} =$$

Fuzzy Set Model for IR - Example

Example :

- Let V = $\{k_1, k_2, k_3, k_4, k_5, k_6\}$

  A = $\{k_1, k_2, k_3\}$ B= $\{k_2, k_3, k_4\}$

- Assume $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

  $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

$$\mu_{A \cup B}(u) = \{k_1:0.8, k_2:0.7, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

$$\mu_{A \cap B}(u) = \{k_1:0, k_2:0.6, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_{\bar{A}} =$$

$$\mu_{\bar{B}} =$$

Fuzzy Set Model for IR - Example

Example :

- Let V = $\{k_1, k_2, k_3, k_4, k_5, k_6\}$

  A = $\{k_1, k_2, k_3\}$ B= $\{k_2, k_3, k_4\}$

- Assume $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

  $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

$$\mu_{A \cup B}(u) = \{k_1:0.8, k_2:0.7, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

$$\mu_{A \cap B}(u) = \{k_1:0, k_2:0.6, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_{\bar{A}} = \{k_1:0.2, k_2:0.3, k_3:0.4, k_4:1, k_5:1, k_6:1\}$$

$$\mu_{\bar{B}} =$$

Fuzzy Set Model for IR - Example

Example :

- Let $V = \{k_1, k_2, k_3, k_4, k_5, k_6\}$

  $A = \{k_1, k_2, k_3\}$ $B = \{k_2, k_3, k_4\}$

- Assume $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

  $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

$$\mu_{A \cup B}(u) = \{k_1:0.8, k_2:0.7, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

$$\mu_{A \cap B}(u) = \{k_1:0, k_2:0.6, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_{\bar{A}} = \{k_1:0.2, k_2:0.3, k_3:0.4, k_4:1, k_5:1, k_6:1\}$$

$$\mu_{\bar{B}} = \{k_1:1, k_2:0.4, k_3:0.2, k_4:0.1, k_5:1, k_6:1\}$$

Fuzzy Set Model for IR

- **Example 2:**
  - Let document A = $\{k_1, k_2, k_3\}$ & document B= $\{k_2, k_3, k_4\}$

$$\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

Query $Q_1 = k_2$ AND k_3

Fuzzy Set Model for IR

- **Example 2:**
  - Let document A = $\{k_1, k_2, k_3\}$ & document B= $\{k_2, k_3, k_4\}$

$$\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

Query $Q_1 = k_2$ AND k_3

$$RSV_A = \min(0.7,\ 0.6) = 0.6$$

$$RSV_B = \min(0.6,\ 0.8) = 0.6$$

Fuzzy Set Model for IR

- **Example 2:**
  - Let document A = $\{k_1, k_2, k_3\}$ & document B= $\{k_2, k_3, k_4\}$

$$\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$$

$$\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$$

Query $Q_1 = k_2$ AND k_3

$$RSV_A = \min(0.7,\ 0.6) = 0.6$$

$$RSV_B = \min(0.6,\ 0.8) = 0.6$$

- **A is ranked the same as B in the result set.**

Fuzzy Set Model for IR

- **Example 3:**
  - Let A = $\{k_1, k_2, k_3\}$ B= $\{k_2, k_3, k_4\}$

    $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

    $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

Find $Q_2 = k_1$ AND k_2 OR k_3

$RSV_A =$

$RSV_B =$

Fuzzy Set Model for IR

- **Example 3:**
  - Let A = $\{k_1, k_2, k_3\}$ B= $\{k_2, k_3, k_4\}$

    $\mu_A = \{k_1:0.8, k_2:0.7, k_3:0.6, k_4:0, k_5:0, k_6:0\}$

    $\mu_B = \{k_1:0, k_2:0.6, k_3:0.8, k_4:0.9, k_5:0, k_6:0\}$

Find $Q_2 = k_1$ AND k_2 OR k_3

$RSV_A = \min(0.8, \max(0.7, 0.6)) = 0.7$

$RSV_B = \min(0, \max(0.6, 0.8)) = 0$

- **A is ranked above B in the result set.**

Ogawa-Morita-Kobayashi Fuzzy IR Model

Fuzzy Information Retrieval (contd.)

- Keyword connection matrix
  - Captures the relationship values that represent the conceptual similarity between two keywords.
  - Gives the normalized co-occurrence of two terms k_i and k_l in a given document d_j.

Y. Ogawa, T. Morita and K. Kobayashi, "Fuzzy document retrieval system and its learning method based on the keyword connection" in: Proc. Int. Workshop on Fuzzy System Applications (1988) 143-144.

Fuzzy Information Retrieval (contd.)

- **Step 1: Construct keyword connection matrix –**
  - Normalized correlation factor $c_{i,l}$ between two terms k_i and k_l (0~1) =

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

  - where $n_{i,l}$ - number of documents containing both term k_i and k_l
    n_i - number of documents containing term k_i
    n_j - number of documents containing term k_l

Fuzzy Information Retrieval (contd.)

- **Step 2: Define a fuzzy set for each term k_i**
  - In the fuzzy set associated to each index term k_i, a document d_j has a degree of membership $\mu_{i,j}$ =

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$$

Fuzzy Information Retrieval (contd.)

- **Step 3: Determine the ranking of documents for the given query using the degree of membership of each document**

- **Step 4: Generate ranked list.**

Class Exercise

For the given corpus, generate the fuzzy ranking of the documents w.r.t the query Q.

- d1: "Shipment of gold damaged in a fire"
  d2: "Delivery of silver arrived in a silver truck"
  d3: "Shipment of gold arrived in a truck"

- q: "gold silver truck"

Class Exercise (contd.)

For the given corpus, generate the fuzzy ranking of the documents w.r.t the query Q.

- d1: "Shipment of gold damaged in a fire"
  d2: "Delivery of silver arrived in a silver truck"
  d3: "Shipment of gold arrived in a truck"

- q: "gold silver truck"

→ **Construct the keyword connection matrix as defined by the Ogawa-Morita-Kobayashi Model**

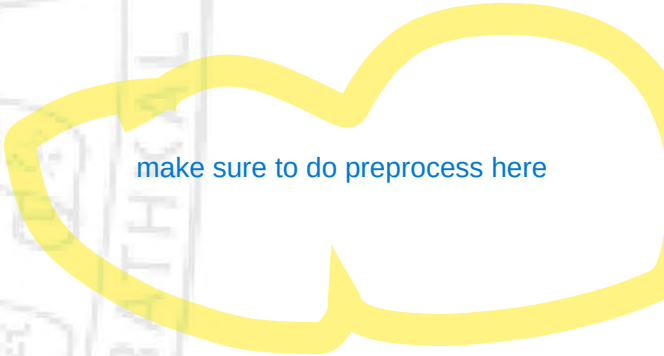
Class Exercise (contd.)

- q: "gold silver truck"
  d1: "Shipment of gold damaged in a fire"
  d2: "Delivery of silver arrived in a silver truck"
  d3: "Shipment of gold arrived in a truck"

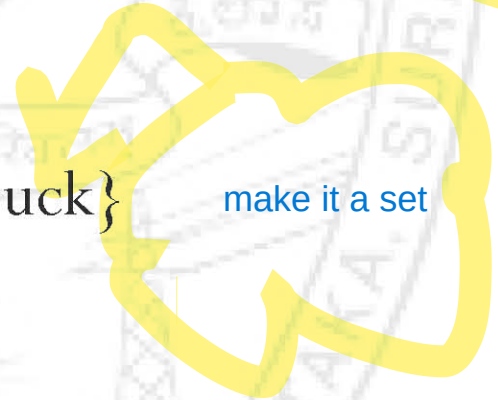
- d1: {shipment, gold, damage, fire}
- d2: {delivery, silver, arrive, silver, truck}
- d3: {shipment, gold, arrive, truck}

Class Exercise (contd.)

- q: "gold silver truck"
  d1: "Shipment of gold damaged in a fire"
  d2: "Delivery of silver arrived in a silver truck"
  d3: "Shipment of gold arrived in a truck"



- d1: {shipment, gold, damage, fire}
- d2: {delivery, silver, arrive, silver, truck}
- d3: {shipment, gold, arrive, truck}



- 8 Keywords (Dimensions) are selected
  - arrive(1), damage(2), delivery(3), fire(4), gold(5), silver(6), shipment(7), truck(8)

Class Exercise (contd.)

- Process:
  - Compute degree of membership of **each query term** w.r.t **each document.**
    - i.e.

$q_{t1}, q_{t2}, q_{t3} \Leftrightarrow d_1 \rightarrow$ (gold, doc1), (silver, doc1), (truck,doc1)

$q_{t1}, q_{t2}, q_{t3} \Leftrightarrow d_2 \rightarrow$ (gold, doc2), (silver, doc2), (truck,doc2)

$q_{t1}, q_{t2}, q_{t3} \Leftrightarrow d_3 \rightarrow$ (gold, doc3), (silver, doc3), (truck,doc3)

  - Process the query w.r.t fuzzy set principles, and find the ranking

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "gold" w.r.t **document d_1**

$$\mu_{gold,d1} = 1 - \prod_{k_1 \in d_1} (1 - C_{gold,k_1})$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "gold" w.r.t **document d_1**

$$\mu_{gold,d1} = 1 - \prod_{k_1 \in d_1} (1 - C_{gold,k_1})$$

$$= 1 - (1 - C_{gold,shipment}) * (1 - C_{gold,gold}) * (1 - C_{gold,damaged}) * (1 - C_{gold,fire})$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "gold" w.r.t **document d_1**

$$\mu_{gold,d1} = 1 - \prod_{k_1 \in d_1} (1 - C_{gold,k_1})$$

$$= 1 - (1 - C_{gold,shipment}) * (1 - C_{gold,gold}) * (1 - C_{gold,damaged}) * (1 - C_{gold,fire})$$

$$= 1 - (1 - \frac{2}{2+2-2}) * (1 - \frac{1}{2+1-1}) * (1 - \frac{2}{2+2-2}) * (1 - \frac{2}{2+1-1})$$

$$= 1 - 0 * \frac{1}{2} * 0 * \frac{1}{2}$$

$$= 1$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "silver" w.r.t **document d_1**

$$\mu_{silver,d1} = 1 - 1 * 1 * 1 * 1 = 0$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "silver" w.r.t **document d_1**

$$\mu_{silver,d1} = 1 - 1 * 1 * 1 * 1 = 0$$

Degree of membership of **query term** "truck" w.r.t **document d_1**

$$\mu_{truck,d1} = 1 - \prod_{k_1 \in d_1} (1 - C_{truck,k_1})$$

$$= 1 - (1 - C_{truck,shipment}) * (1 - C_{truck,gold}) * (1 - C_{truck,damaged}) * (1 - C_{truck,fire})$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "silver" w.r.t **document d_1**

$$\mu_{silver,d1} = 1 - 1 * 1 * 1 * 1 = 0$$

Degree of membership of **query term** "truck" w.r.t **document d_1**

$$\mu_{truck,d1} = \frac{5}{9}$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "gold","silver" and "truck" w.r.t **document d_2**

$$\mu_{gold,d2} = 1 - 1 * 1 * \frac{2}{3} * \frac{2}{3} = \frac{5}{9}$$

$$\mu_{silver,d2} = 1$$

$$\mu_{truck,d2} = 1$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Degree of membership of **query term** "gold", "silver" and "truck" w.r.t **document d_3**

$$\mu_{gold,d3} = 1$$

$$\mu_{silver,d3} = 1 - 1 * 1 * \frac{1}{2} * \frac{1}{2} = \frac{3}{4}$$

$$\mu_{truck,d3} = 1$$

Class Exercise (contd.)

d1:"Shipment of gold damaged in a fire"
d2:"Delivery of silver arrived in a silver truck"
d3:"Shipment of gold arrived in a truck"

q: "gold silver truck"

Compute similarity of Query q with each document

$$\mu_{q,d1} = \mu_{gold \wedge silver \wedge truck,d1} = \min(\mu_{gold,d1}, \mu_{silver,d1}, \mu_{truck,d1}) = 0$$

$$\mu_{q,d2} = \mu_{gold \wedge silver \wedge truck,d2} = \min(\mu_{gold,d2}, \mu_{silver,d2}, \mu_{truck,d2}) = \frac{5}{9}$$

$$\mu_{q,d3} = \mu_{gold \wedge silver \wedge truck,d3} = \min(\mu_{gold,d3}, \mu_{silver,d3}, \mu_{truck,d3}) = \frac{3}{4}$$

Class Exercise (contd.)

Compute similarity of Query q with each document

$$\mu_{q,d1} = \mu_{gold \wedge silver \wedge truck,d1} = \min(\mu_{gold,d1}, \mu_{silver,d1}, \mu_{truck,d1}) = 0$$

$$\mu_{q,d2} = \mu_{gold \wedge silver \wedge truck,d2} = \min(\mu_{gold,d2}, \mu_{silver,d2}, \mu_{truck,d2}) = \frac{5}{9}$$

$$\mu_{q,d3} = \mu_{gold \wedge silver \wedge truck,d3} = \min(\mu_{gold,d3}, \mu_{silver,d3}, \mu_{truck,d3}) = \frac{3}{4}$$

$Sim(q,d_3) > Sim(q,d_2) > Sim(q,d_1)$

Fuzzy IR – Summary

- Advantages –
  - Correlations among index terms are considered
  - Degree of relevance between queries and docs can be achieved
    - Ranked retrieval
- Disadvantages –
  - Does not consider the frequency of a term in a document or a query.
  - Relevance feedback not attempted.

Further reading...

- Tahani, Valiollah. "A fuzzy model for document retrieval systems." *Information Processing & Management* 12.3 (1976): 177-187.

- Y. Ogawa, T. Morita and K. Kobayashi, "Fuzzy document retrieval system and its learning method based on the keyword connection" in: Proc. Int. Workshop on Fuzzy System Applications (1988) 143-144.