

The Fuzzy Set Theory

The **Fuzzy Set Theory** deals with the representation of classes whose boundaries are not well defined

A fuzzy subset of a u.o.d. U is characterized by a membership function $\mu_A : U \rightarrow [0,1]$ which associates with each element $u \in U$ a number $\mu_A(u)$ in the interval $[0,1]$

- 0 means no membership
- 1 means full membership
- $0 < x < 1$ something in between

Commonly used operations on fuzzy sets are:

- the complement $\mu_{\bar{A}}(u) = 1 - \mu_A(u)$
- the union $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
- the intersection $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

The Fuzzy Logic Model (FLM) [Tahani,1976]

The Fuzzy Logic Model (FLM), proposed by Tahani (1976) is an extension, in a quantitative direction, of the boolean model:

- A document is represented by an and of weighted index terms;
e.g. $d_i = \langle \text{and } \langle \text{jazz}, 0.4 \rangle \langle \text{sax}, 0.3 \rangle \langle \text{Coltrane}, 0.5 \rangle \rangle$;
- A query q_j is represented as in the Boolean Model
- The matching function RSV computes the degree to which document d_i satisfies q_j
 - $RSV(d_i, t_k) = w_{ki}$
 - $RSV(d_i, (\text{not } q_j)) = 1 - RSV(d_i, q_j)$
 - $RSV(d_i, (\text{and } q_1 \dots q_n)) = \min\{RSV(d_i, q_1), \dots, RSV(d_i, q_n)\}$
 - $RSV(d_i, (\text{or } q_1 \dots q_n)) = \max\{RSV(d_i, q_1), \dots, RSV(d_i, q_n)\}$

Here too there is an implicit **closed word assumption**: the absence of a term t in the representation of d is only a notational abbreviation for the presence of $\langle t, 0 \rangle$ in the same representation

Ogawa-Morita-Kobayashi implementation [1991]

- The basic idea is to expand the set of index terms in the query with related terms such that additional relevant documents can be retrieved by the user query
- A thesaurus is constructed by defining a term-term correlation matrix (**keyword connection matrix**)

$$c_{il} = \frac{n_{il}}{n_i + n_l - n_{il}}$$

n_i : #docs containing term k_i

n_l : #docs containing term k_l

n_{il} : #docs containing both of them

- We can use this matrix to define a fuzzy set associated to each index term k_i

- In this fuzzy set, a document d_j has degree of membership

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

- Given a query as a DNF $q = cc_1 \vee cc_2 \dots \vee cc_p$

$$\mu_{qj} = \mu_{(\sum_i cc_i)j} = 1 - \prod_{i=1}^p (1 - \mu_{cc_i j}) = 1 - \prod_i (1 - \prod_k \mu_{cc_i^k j})$$

N.B. We have used **algebraic sum** (implemented as the **complement of negated algebraic product**) in place of the **max** function, and **algebraic product** in place of the **min** function, respectively.

Advantages of FLM

1. To a user familiar with Boolean systems, a fuzzy system is different only in that it returns ranked output
2. If the weights $w_{ki} \in \{0,1\}$, the FLM behaves like the BM; a FLM-based IR system is thus compatible also with document representations produced for Boolean systems
3. Possibility to associate weights to index terms in the IREPs of documents
4. Output magnitude control:
 - No more under-dimensioned output, as the terms that represent a document will be many more than in a Boolean representation, even if they have weights 0. As a consequence, almost always at least one document satisfies a query with degree >0
 - No more over-dimensioned output as, thanks to ranked retrieval, the user may scan the ranked list until desired

Disadvantages of FLM

1. As in the BM, no automatic query acquisition;
2. As in the BM, no possibility to associate weights to index terms in the IREPs of queries. Extensions of the FLM to allow this are mathematically complicated
3. As in the BM, flattened, hence unintuitive, results (lack of output discrimination)! The problem derives from the fact that documents are ranked based only on the weight of a single term
 - Following query (and jazz flute Coltrane) the same RSV of 0.1 is given to document
(and $\langle \text{jazz}, 0.1 \rangle \langle \text{flute}, 0.1 \rangle \langle \text{Coltrane}, 0.1 \rangle \langle \text{Dolphy}, 0.4 \rangle$) and to document
(and $\langle \text{jazz}, 0.1 \rangle \langle \text{flute}, 0.9 \rangle \langle \text{Coltrane}, 0.9 \rangle \langle \text{Parker}, 0.3 \rangle$) alike;
 - Following query (or jazz flute Coltrane Dolphy) the same RSV of 0.8 is given to document
(and $\langle \text{jazz}, 0.5 \rangle \langle \text{flute}, 0.5 \rangle \langle \text{Coltrane}, 0.5 \rangle \langle \text{Dolphy}, 0.8 \rangle$) and to document
(and $\langle \text{classical}, 0.1 \rangle \langle \text{flute}, 0.8 \rangle \langle \text{Gazzelloni}, 0.1 \rangle$) alike.