Models for Retrieval and Browsing

- Fuzzy Set, Extended Boolean, Generalized Vector Space Models

Berlin Chen 2003

Reference:

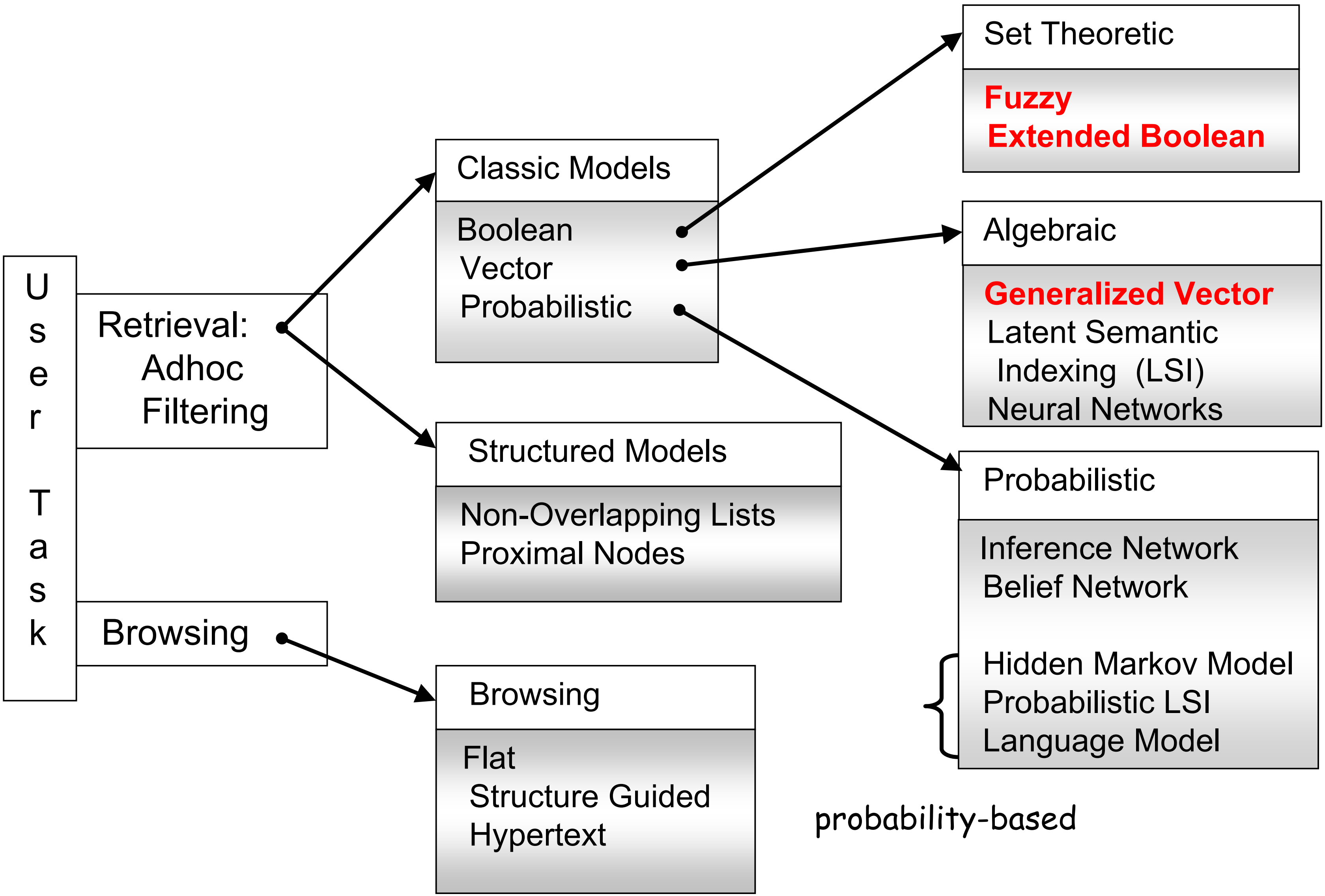
1. Modern Information Retrieval, chapter 2

Outline

- Alternative Set Theoretic Models
  - Fuzzy Set Model (Fuzzy Information Retrieval)
  - Extended Boolean Model

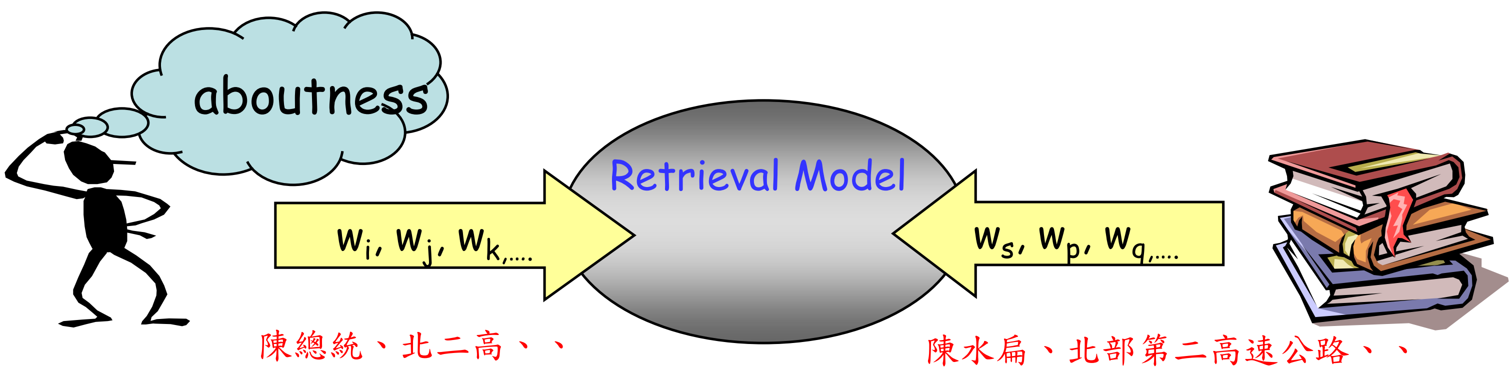
- Alternative Algebraic Models
  - Generalized Vector Space Model

Taxonomy of Classic IR Models



Fuzzy Set Model

- Premises
  - Docs and queries are represented through sets of keywords, therefore the matching between them is vague
    - Keywords cannot completely describe the user's information need and the doc's main theme



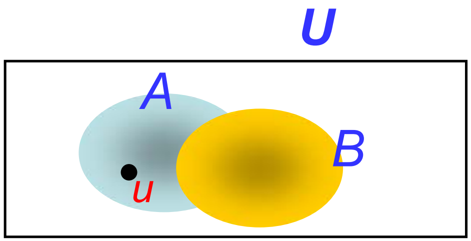
  - For each query term (keyword)
    - Define a fuzzy set and that each doc has a degree of membership (0~1) in the set

Fuzzy Set Model

- Fuzzy Set Theory
  - Framework for representing classes (sets) whose boundaries are not well defined
  - Key idea is to introduce the notion of a *degree of membership* associated with the elements of a set
  - This degree of membership varies from 0 to 1 and allows modeling the notion of *marginal* membership
    - 0 → no membership
    - 1 → full membership
  - Thus, membership is now a gradual instead of abrupt
    - Not as conventional Boolean logic

Here we will define a fuzzy set for each query (or index) term, thus each doc has a degree of membership in this set.

Fuzzy Set Model



- Definition
  - A fuzzy subset A of a universal of discourse U is characterized by a membership function $\mu_A: U \rightarrow [0,1]$
    - Which associates with each element u of U a number $\mu_A(u)$ in the interval [0,1]
  - Let A and B be two fuzzy subsets of U. Also, let $\overline{A}$ be the complement of A. Then,
    - Complement $\mu_{\overline{A}}(u) = 1 - \mu_A(u)$
    - Union $\mu_{A \cup B}(u) = \max(\mu_A(u), \mu_B(u))$
    - Intersection $\mu_{A \cap B}(u) = \min(\mu_A(u), \mu_B(u))$

Fuzzy Set Model

- Fuzzy information retrieval

Defining term relationship

  - Fuzzy sets are modeled based on a **thesaurus**
  - This thesaurus can be constructed by a term-term correlation matrix (or called keyword connection matrix)
    - $\vec{c}$: a term-term correlation matrix
    - $c_{i,l}$: a normalized correlation factor for terms k_i and k_l

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

ranged from 0 to 1

n_i : no of docs that contain k_i

$n_{i,l}$: no of docs that contain both k_i and k_l

docs, paragraphs, sentences, ..

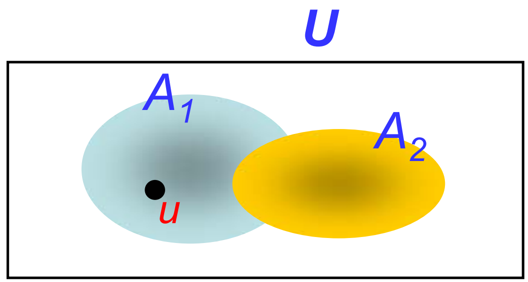
    - We now have the notion of proximity among index terms

  - The relationship is symmetric !

$$\mu_{k_i}(k_l) = c_{i,l} = c_{l,i} = \mu_{k_l}(k_i)$$

Fuzzy Set Model

- The union and intersection operations are modified here



$$
\begin{aligned}
&ab + \bar{a}b + a\bar{b} \\
&= ab + (1-a)b + a(1-b) \\
&= ab + b - ab + a - ab \\
&= 1 - (1 - a - b + ab) \\
&= 1 - (1-a)(1-b)
\end{aligned}
$$

- Union: algebraic sum (instead of max)

$$\mu_{A_1 \cup A_2}(u) = \mu_{A_1}(u)\mu_{A_2}(u) + \mu_{\overline{A_1}}(u)\mu_{A_2}(u) + \mu_{A_1}(u)\mu_{\overline{A_2}}(u)$$

$$= 1\text{-}\prod_{j=1}^{2}\left(1\text{-}\mu_{A_j}(u)\right)$$

a negative algebraic product

$$\Rightarrow \quad \mu_{A_1 \cup A_2 \cdots \cup A_n}(u) = \mu_{\cup_j A_j}(u) = 1\text{-}\prod_{j=1}^{n}\left(1\text{-}\mu_{A_j}(u)\right)$$

- Intersection: algebraic product (instead of min)

$$\mu_{A_1 \cap A_2}(u) = \mu_{A_1}(u)\mu_{A_2}(u) \quad \Rightarrow \quad \mu_{A_1 \cap A_2 \cdots \cap A_n}(u) = \prod_{j=1}^{n}\mu_{A_j}(u)$$

Fuzzy Set Model

– The degree of membership between a doc d_j and an index term k_i

algebraic sum (a doc is a union of index terms)

$$\mu_{k_i}(d_j) = \mu_{d_j}(k_i) = \mu_{\underset{k_l \in d_j}{\cup} k_l}(k_i)$$

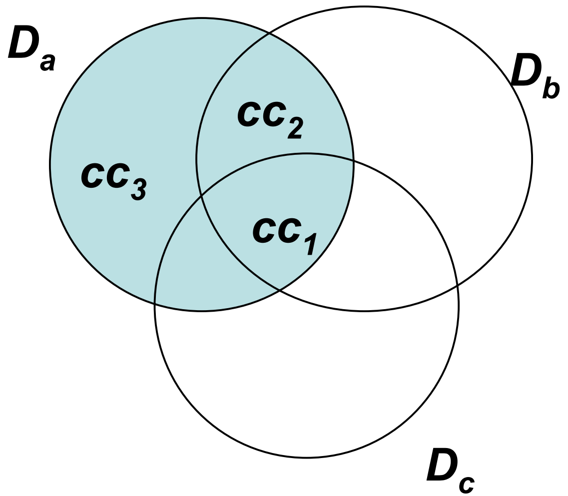
$$= 1 - \prod_{k_l \in d_j}\left(1 - \mu_{k_l}(k_i)\right) = 1 - \prod_{k_l \in d_j}\left(1 - c_{i,l}\right)$$

	k_a	k_b
k_i	$c_{i,a}$	$c_{i,b}$
	$1 - c_{i,a}$	$1 - c_{i,b}$

- Computes an **algebraic sum** over all terms in the doc d_j
  - Implemented as the complement of a negative algebraic product
  - A doc d_j belongs to the fuzzy set associated to the term k_i if its own terms are related to k_i
- If there is at least one index term k_l of d_j which is strongly related to the index k_i ($c_{i,l} \sim 1$) then $\mu_{k_i,d_j} \sim 1$
  - k_i is a good fuzzy index for doc d_j
  - And vice versa

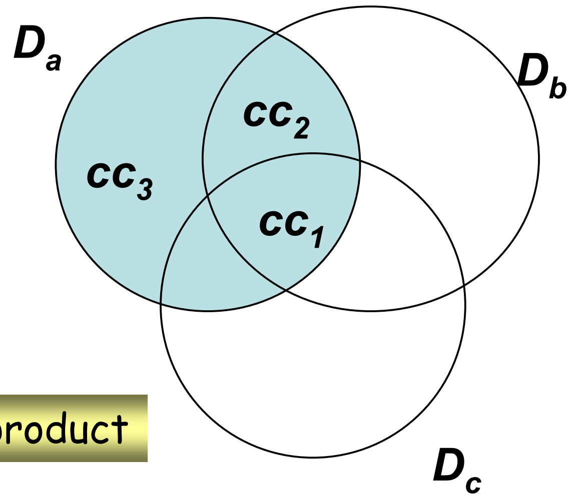
Fuzzy Set Model

- Example:
  - Query $q = k_a \wedge (k_b \vee \neg k_c)$ disjunctive normal form
    $\vec{q}_{dnf} = (k_a \wedge k_b \wedge k_c) \vee (k_a \wedge k_b \wedge \neg k_c) \vee (k_a \wedge \neg k_b \wedge \neg k_c)$
    $= cc_1 + cc_2 + cc_3$ ← conjunctive component



  - D_a is the fuzzy set of docs associated to the term k_a
  - Degree of membership ?

Fuzzy Set Model



- Degree of membership

algebraic sum

$$\mu_{q,d_j} = \mu_{cc_1 \cup cc_2 \cup cc_3, d_j}$$

for a doc d_j in the fuzzy answer set D_q

negative algebraic product

$$= 1 - \prod_{i=1}^{3} (1 - \mu_{cc_i, d_j})$$

$$= 1 - \Big(1 - \underbrace{\mu_{a \cap b \cap c, d_j}}_{cc_1}\Big)\Big(1 - \underbrace{\mu_{a \cap b \cap \bar{c}, d_j}}_{cc_2}\Big)\Big(1 - \underbrace{\mu_{a \cap \bar{b} \cap \bar{c}, d_j}}_{cc_3}\Big)$$

algebraic product

$$= 1 - (1 - \mu_{a,d_j} \mu_{b,d_j} \mu_{c,d_j})$$

$$\times (1 - \mu_{a,d_j} \mu_{b,d_j} (1 - \mu_{c,d_j})) \times (1 - \mu_{a,d_j} (1 - \mu_{b,d_j})(1 - \mu_{c,d_j}))$$

Fuzzy Set Model

- Advantages
  - The correlations among index terms are considered
  - Degree of relevance between queries and docs can be achieved
- Disadvantages
  - Fuzzy IR models have been discussed mainly in the literature associated with fuzzy theory
  - Experiments with standard test collections are not available

Fuzzy Set Model

- Advantages
  - The correlations between index terms are considered
- Disadvantages
  - Fuzzy IR models have been discussed mainly in the literature associated with fuzzy theory
  - Experiments with standard test collections are not available

Extended Boolean Model

Salton et al., 1983

- Motive
  - Extend the Boolean model with the functionality of partial matching and term weighting 陳水扁 及 呂秀蓮
    - E.g.: in Boolean model, for the qery $q=k_x \wedge k_y$, a doc contains either k_x or k_y is as irrelevant as another doc which contains neither of them
    - How about the disjunctive query $q=k_x \vee k_y$ 陳水扁 或 呂秀蓮
  - Combine Boolean query formulations with characteristics of the vector model
    - Term weighting
    - Algebraic distances for similarity measures

} a ranking can be obtained

Extended Boolean Model

- Term weighting
  - The weight for the term k_x in a doc d_j is

$$w_{x,j} = tf_{x,j} \times \frac{idf_x}{\max_i idf_i}$$

ranged from 0 to 1

normalized frequency

Normalized *idf*

    - $w_{x,j}$ is normalized to lay between 0 and 1

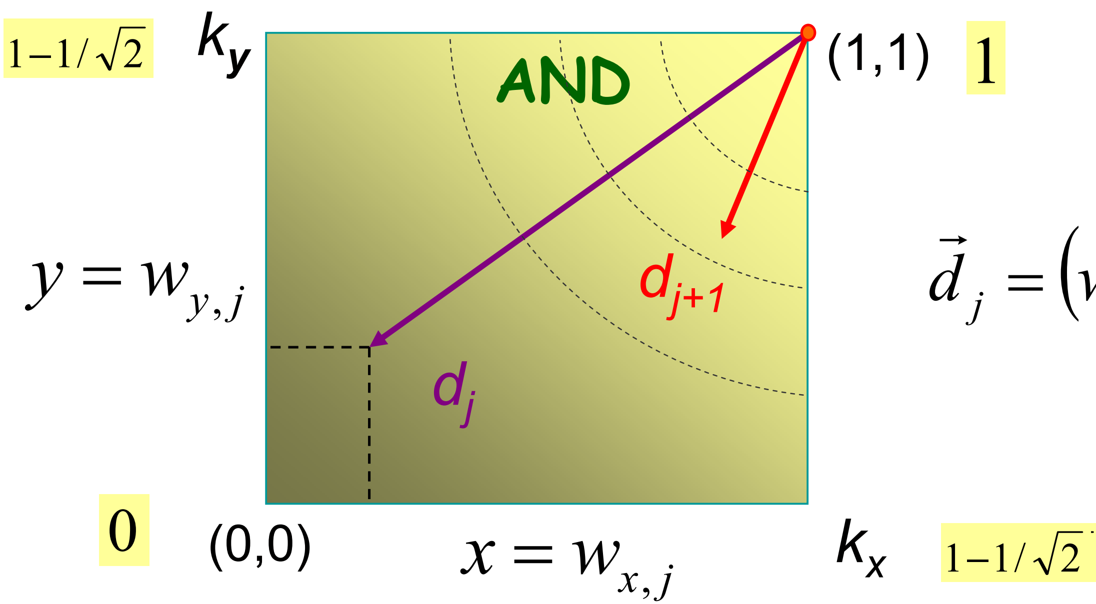
- Assume two index terms k_x and k_y were used
  - Let x denote the weight $w_{x,j}$ of term k_x on doc d_j
  - Let y denote the weight $w_{y,j}$ of term k_y on doc d_j
  - The doc vector $\vec{d}_j = (w_{x,j}, w_{y,j})$ is represented as $d_j = (x, y)$
  - Queries and docs can be plotted in a two-dimensional map

Extended Boolean Model

- If the query is $q = k_x \wedge k_y$ (conjunctive query)
  - -The docs near the point (1,1) are preferred
  - -The similarity measure is defined as

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

2-norm model
(Euclidean distance)



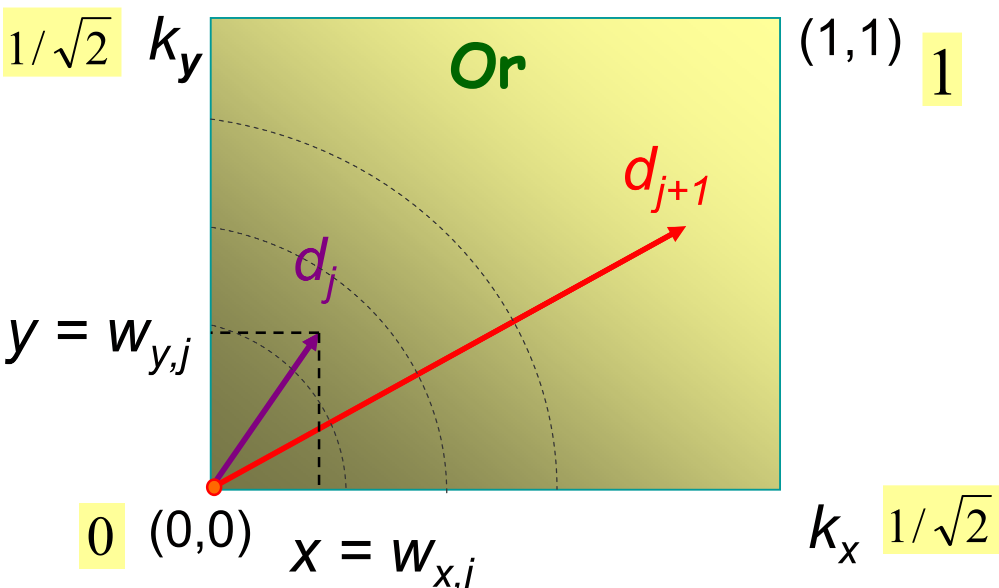
$$\vec{d}_j = (w_{x,j}, w_{y,j})$$

Extended Boolean Model

- If the query is $q=k_x \vee k_y$ (disjunctive query)
  - -The docs far from the point (0,0) are preferred
  - -The similarity measure is defined as

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

2-norm model
(Euclidean distance)



Extended Boolean Model

- The similarity measures $sim\left(q_{or}, d\right)$ and $sim\left(q_{and}, d\right)$ also lay between 0 and 1

Extended Boolean Model

- Generalization
  - *t* index terms are used $\rightarrow$ *t*-dimensional space
  - *p*-norm model, $1 \leq p \leq \infty$

$$q_{and} = k_1 \wedge^p k_2 \wedge^p \ldots \wedge^p k_m \Rightarrow sim(q_{and}, d) = 1 - \left(\frac{(1-x_1)^p + (1-x_2)^p + \ldots + (1-x_m)^p}{m} \right)^{\frac{1}{p}}$$

$$q_{or} = k_1 \vee^p k_2 \vee^p \ldots v^p k_m \Rightarrow sim(q_{or}, d) = \left(\frac{x_1^{\ p} + x_2^{\ p} + \ldots + x_m^{\ p}}{m} \right)^{\frac{1}{p}}$$

  - Some interesting properties
    - *p=1* $\Rightarrow sim(q_{and}, d) = sim(q_{or}, d) = \frac{x_1 + x_2 + \ldots + x_m}{m}$
    - *p=* $\infty \Rightarrow sim(q_{and}, d) \approx \min(x_i)$

      $sim(q_{or}, d) \approx \max(x_i)$

just like the formula of fuzzy logic

Extended Boolean Model

- Example query 1: $q = (k_1 \wedge^p k_2) \vee^p k_3$
  - Processed by grouping the operators in a predefined order

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

- Example query 2: $q = (k_1 \vee^2 k_2) \wedge^\infty k_3$
  - Combination of different algebraic distances

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

Extended Boolean Model

- Advantages
  - A hybrid model including properties of both the set theoretic models and the algebraic models
    - Relax the Boolean algebra by interpreting Boolean operations in terms of algebraic distances
- Disadvantages
  - Distributive operation does not hold for ranking computation
    - E.g.: $q_1 = \left(k_1 \wedge^2 k_2\right) \vee^2 k_3, q_2 = \left(k_1 \vee^2 k_3\right) \wedge^2 \left(k_2 \vee^2 k_3\right)$

$$\left(\frac{\left(1 - \left(\frac{(1-x_1)^2 + (1-x_2)^2}{2}\right)^{\frac{1}{2}}\right)^2 + x_3^2}{2}\right)^{\frac{1}{2}} \qquad sim\left(q_1, d\right) \neq sim\left(q_2, d\right) \qquad 1 - \left(\frac{\left(1 - \left(\frac{x_1^2 + x_2^2}{2}\right)\right)^2 + \left(1 - \left(\frac{x_2^2 + x_3^2}{2}\right)\right)^2}{2}\right)^{\frac{1}{2}}$$

  - Assumes mutual independence of index terms

Generalized Vector Model

Wong et al., 1985

- Premise
  - Classic models enforce *independence* of index terms
  - For the **Vector model**
    - Set of term vectors $\{\vec{k}_1, \vec{k}_1, \ldots, \vec{k}_t\}$ are linearly independent and form a basis for the subspace of interest
    - Frequently, it means *pairwise orthogonality*

      $\forall i,j \Rightarrow \vec{k}_i \bullet \vec{k}_j = \vec{0}$ *(in a more restrictive sense)*

- Wong et al. proposed an interpretation
  - The index term vectors are *linearly independent*, but *not pairwise orthogonal*
    - Generalized Vector Model

Generalized Vector Model

- **Key idea**
  - Index term vectors form the basis of the space are not orthogonal and are represented in terms of smaller components (minterms)

- **Notations**
  - $\{k_1, k_2, ..., k_t\}$: the set of all terms
  - $w_{i,j}$: the weight associated with $[k_i, d_j]$
  - **Minterms**:binary indicators (0 or 1) of all patterns of occurrence of terms within documents
    - Each represent one kind of co-occurrence of index terms in a specific document

Generalized Vector Model

- Representations of **minterms**

$m_1=(0,0,\ldots,0)$
$m_2=(1,0,\ldots,0)$
$m_3=(0,1,\ldots,0)$
$m_4=(1,1,\ldots,0)$
$m_5=(0,0,1,..,0)$
…
$m_{2^t}=(1,1,1,..,1)$

2^t minterms

Points to the docs where only index terms k_1 and k_2 co-occur and the other index terms disappear

Point to the docs containing all the index terms

$\vec{m}_1=(1,0,0,0,0,\ldots,0)$
$\vec{m}_2=(0,1,0,0,0,\ldots,0)$
$\vec{m}_3=(0,0,1,0,0,\ldots,0)$
$\vec{m}_4=(0,0,0,1,0,\ldots,0)$
$\vec{m}_5=(0,0,0,0,1,\ldots,0)$
…
$\vec{m}_{2^t}=(0,0,0,0,0,\ldots,1)$

2^t minterm vectors

Pairwise orthogonal vectors $\vec{m}_i$ associated with minterms m_i as the **basis** for the ***generalized vector space***

Generalized Vector Model

- Minterm vectors are pairwise orthogonal. But, this does not mean that the index terms are independent
  - Each minterm specifies a kind of dependence among index terms
  - That is, the co-occurrence of index terms inside docs in the collection induces dependencies among these index terms

Generalized Vector Model

- The vector associated with the term k_i is represented by **summing** up all minterms containing it and **normalizing**

$$\vec{k}_i = \frac{\sum_{\forall r, g_i(m_r)=1} c_{i,r} \vec{m}_r}{\sqrt{\sum_{\forall r, g_i(m_r)=1} c_{i,r}^2}}$$

$$c_{i,r} = \sum_{d_j \mid g_l(\vec{d}_j) = g_l(m_r), \text{ for all } l} w_{i,j}$$

- The weight associated with the pair $[k_i, m_r]$ sums up the weights of the term k_i in all the docs which have a term occurrence pattern given by m_r.
- Notice that for a collection of size N, only N minterms affect the ranking (and not 2^N)

All the docs whose term co-occurrence relation (pattern) can be represented as (exactly coincide with that of) minterm m_r

$g_i(m_r)$ Indicates the index term k_i is in the minterm m_r

Generalized Vector Model

- The similarity between the query and doc is calculated in the space of minterm vectors

$$\vec{d}_j = \sum_i w_{i,j} \vec{k}_i \quad \Rightarrow \quad = \sum_r s_{j,r} \vec{m}_r$$

$$\vec{q}_j = \sum_i w_{i,q} \vec{k}_i \quad \Rightarrow \quad = \sum_r s_{q,r} \vec{m}_r$$

t-dimensional $\qquad$ 2^t-dimensional

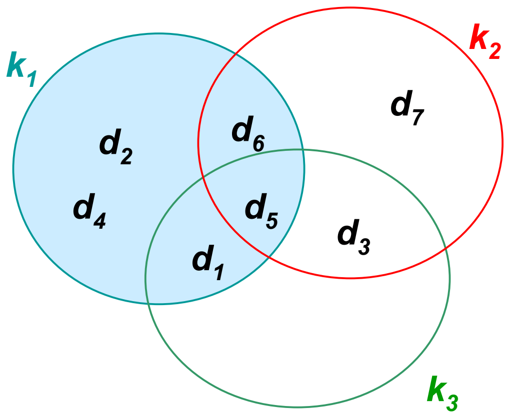
$$sim\left(\vec{q}_j, \vec{d}_j\right) = \frac{\sum_i w_{i,q} \cdot w_{i,j}}{\sqrt{\sum_i w_{i,q}} \sqrt{\sum_i w_{i,q}}}$$

$$sim\left(\vec{q}_j, \vec{d}_j\right) = \frac{\sum_r s_{q,r} \cdot s_{d,r}}{\sqrt{\sum_r s_{q,r}} \sqrt{\sum_r s_{d,r}}}$$

Generalized Vector Model

- **Example** (a system with three index terms)

minterm	k_1	k_2	k_3
m_1	0	0	0
m_2	1	0	0
m_3	0	1	0
m_4	1	1	0
m_5	0	0	1
m_6	1	0	1
m_7	0	1	1
m_8	1	1	1



$$\vec{k}_1 = \frac{c_{1,2}\vec{m}_2 + c_{1,4}\vec{m}_4 + c_{1,6}\vec{m}_6 + c_{1,8}\vec{m}_8}{\sqrt{c_{1,2}^{\;2} + c_{1,4}^{\;2} + c_{1,6}^{\;2} + c_{1,8}^{\;2}}}$$

$$\vec{k}_2 = \frac{c_{2,3}\vec{m}_3 + c_{2,4}\vec{m}_4 + c_{2,7}\vec{m}_7 + c_{2,8}\vec{m}_8}{\sqrt{c_{2,3}^{\;2} + c_{2,4}^{\;2} + c_{2,7}^{\;2} + c_{2,8}^{\;2}}}$$

$$\vec{k}_3 = \frac{c_{3,5}\vec{m}_5 + c_{3,6}\vec{m}_6 + c_{3,7}\vec{m}_7 + c_{3,8}\vec{m}_8}{\sqrt{c_{3,5}^{\;2} + c_{3,6}^{\;2} + c_{3,7}^{\;2} + c_{3,8}^{\;2}}}$$

	k_1	k_2	k_3	minterm
d_1	2	0	1	m_6
d_2	1	0	0	m_2
d_3	0	1	3	m_7
d_4	2	0	0	m_2
d_5	1	2	4	m_8
d_6	1	2	0	m_4
d_7	0	5	0	m_3
q	1	2	3	

$$c_{1,2} = w_{1,2} + w_{1,4} = 1 + 2 = 3$$
$$c_{1,4} = w_{1,6} = 1$$
$$c_{1,6} = w_{1,1} = 2$$
$$c_{1,8} = w_{1,5} = 1$$

$$\vec{k}_1 = \frac{3\vec{m}_2 + 1\vec{m}_4 + 2\vec{m}_6 + 1\vec{m}_8}{\sqrt{3^2 + 1^2 + 2^2 + 1^2}}$$

$$c_{2,3} = w_{2,7} = 5$$
$$c_{2,4} = w_{2,6} = 2$$
$$c_{2,7} = w_{2,3} = 1$$
$$c_{2,8} = w_{2,5} = 2$$

$$\vec{k}_2 = \frac{5\vec{m}_3 + 2\vec{m}_4 + 1\vec{m}_7 + 2\vec{m}_8}{\sqrt{5^2 + 2^2 + 1^2 + 2^2}}$$

$$c_{3,5} = 0$$
$$c_{3,6} = w_{3,1} = 1$$
$$c_{3,7} = w_{3,3} = 3$$
$$c_{3,8} = w_{3,5} = 4$$

$$\vec{k}_3 = \frac{0\vec{m}_5 + 1\vec{m}_6 + 3\vec{m}_7 + 4\vec{m}_8}{\sqrt{0^2 + 1^2 + 3^2 + 4^2}}$$

Generalized Vector Model

- **Example**: Ranking

$$\vec{k}_1 = \frac{3\vec{m}_2 + 1\vec{m}_4 + 2\vec{m}_6 + 1\vec{m}_8}{\sqrt{3^2+1^2+2^2+1^2}} = \frac{3\vec{m}_2 + 1\vec{m}_4 + 2\vec{m}_6 + 1\vec{m}_8}{\sqrt{15}}$$

$$\vec{k}_2 = \frac{5\vec{m}_3 + 2\vec{m}_4 + 1\vec{m}_7 + 2\vec{m}_8}{\sqrt{5^2+2^2+1^2+2^2}} = \frac{5\vec{m}_3 + 2\vec{m}_4 + 1\vec{m}_7 + 2\vec{m}_8}{\sqrt{34}}$$

$$\vec{k}_3 = \frac{0\vec{m}_5 + 1\vec{m}_6 + 3\vec{m}_7 + 4\vec{m}_8}{\sqrt{0^2+1^2+3^2+4^2}} = \frac{1\vec{m}_6 + 3\vec{m}_7 + 4\vec{m}_8}{\sqrt{26}}$$

$$\vec{d}_1 = 2\vec{k}_1 + 1\vec{k}_3$$

$$= \underbrace{\frac{2\cdot 3}{\sqrt{15}}}_{s_{d1,2}}\vec{m}_2 + \underbrace{\frac{2\cdot 1}{\sqrt{15}}}_{s_{d1,4}}\vec{m}_4 + \underbrace{\left(\frac{2\cdot 2}{\sqrt{15}} + \frac{1\cdot 1}{\sqrt{26}}\right)}_{s_{d1,6}}\vec{m}_6 + \underbrace{\frac{1\cdot 3}{\sqrt{26}}}_{s_{d1,7}}\vec{m}_7 + \underbrace{\left(\frac{2\cdot 1}{\sqrt{15}} + \frac{1\cdot 4}{\sqrt{26}}\right)}_{s_{d1,8}}\vec{m}_8$$

$$\vec{q} = 1\vec{k}_1 + 2\vec{k}_2 + 3\vec{k}_3$$

$$= \underbrace{\frac{1\cdot 3}{\sqrt{15}}}_{s_{q,2}}\vec{m}_2 + \underbrace{\frac{2\cdot 5}{\sqrt{34}}}_{s_{q,3}}\vec{m}_3 + \underbrace{\left(\frac{1\cdot 1}{\sqrt{15}} + \frac{2\cdot 2}{\sqrt{34}}\right)}_{s_{q,4}}\vec{m}_4 + \underbrace{\left(\frac{1\cdot 2}{\sqrt{15}} + \frac{3\cdot 1}{\sqrt{26}}\right)}_{s_{q,6}}\vec{m}_6 + \underbrace{\left(\frac{2\cdot 1}{\sqrt{34}} + \frac{3\cdot 3}{\sqrt{26}}\right)}_{s_{q,7}}\vec{m}_7 + \underbrace{\left(\frac{1\cdot 1}{\sqrt{15}} + \frac{2\cdot 2}{\sqrt{34}} + \frac{3\cdot 4}{\sqrt{26}}\right)}_{s_{q,8}}\vec{m}_8$$

$$sim(q,d) = \text{consine}(q,d) = \frac{\sum_{r \mid s_{q,r} \neq 0 \wedge s_{d,r} \neq 0} s_{q,r} \cdot s_{d,r}}{\sqrt{\sum_r s_{q,r}^2}\sqrt{\sum_r s_{d,r}^2}}$$

The similarity between the query and doc is calculated in the space of minterm vectors

$$sim(q,d_1) = \frac{s_{q,2}s_{d_1,2} + s_{q,4}s_{d_1,4} + s_{q,6}s_{d_1,6} + s_{q,7}s_{d_1,7} + s_{q,8}s_{d_1,8}}{\sqrt{s_{q,2}^2 + s_{q,3}^2 + s_{q,4}^2 + s_{q,6}^2 + s_{q,7}^2 + s_{q,8}^2}\sqrt{s_{d_1,2}^2 + s_{d_1,4}^2 + s_{d_1,6}^2 + s_{d_1,7}^2 + s_{d_1,8}^2}}$$

Generalized Vector Model

- Term Correlation
  - The degree of correlation between the terms k_i and k_j can now be computed as

$$\vec{k}_i \bullet \vec{k}_j = \sum_{\forall r | g_i(m_r)=1 \wedge g_j(m_r)=1} c_{i,r} \times c_{j,r}$$

    - Do not need to be normalized? (because we have done it before!)

Generalized Vector Model

- Advantages
  - Model considers correlations among index terms
  - Model does introduce interesting new ideas

- Disadvantages
  - Not clear in which situations it is superior to the standard vector model
  - Computation costs are higher