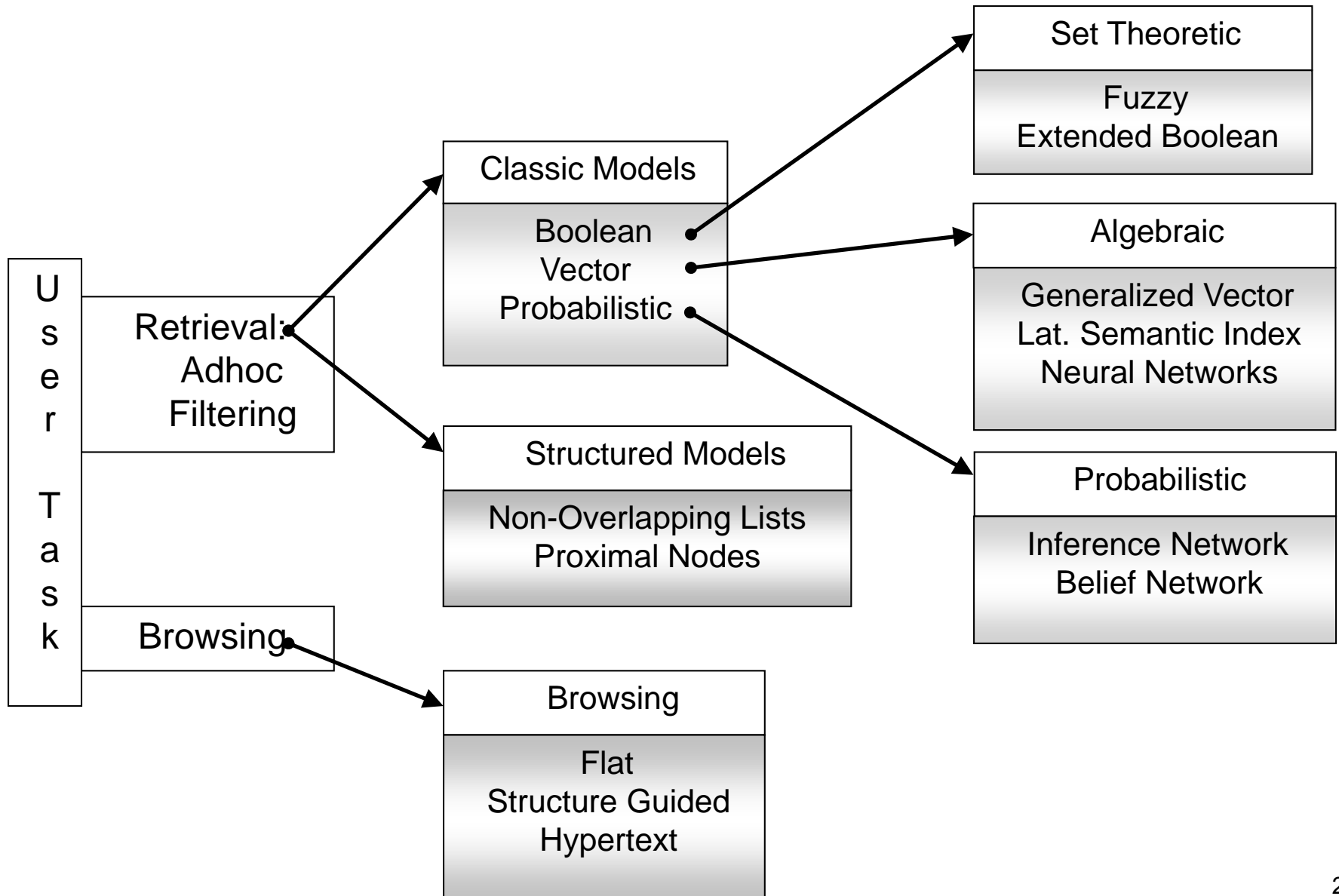

Chapter 02

Modeling

IR Models



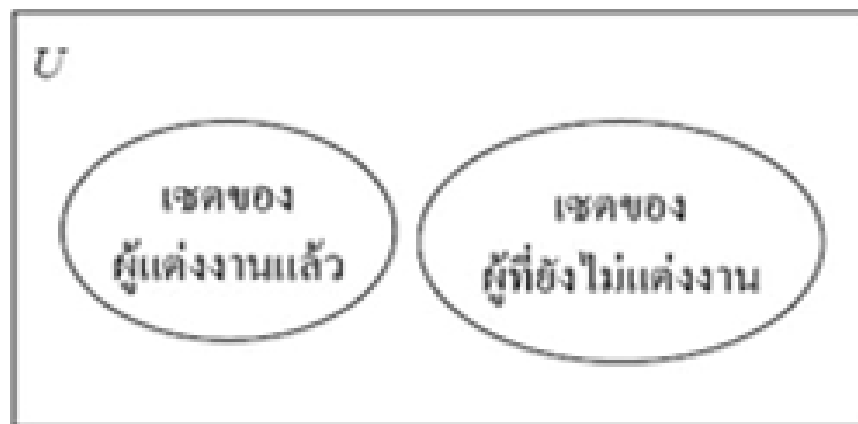
Fuzzy logic

เซตแบบต้นฉบับ (Crisp Set)

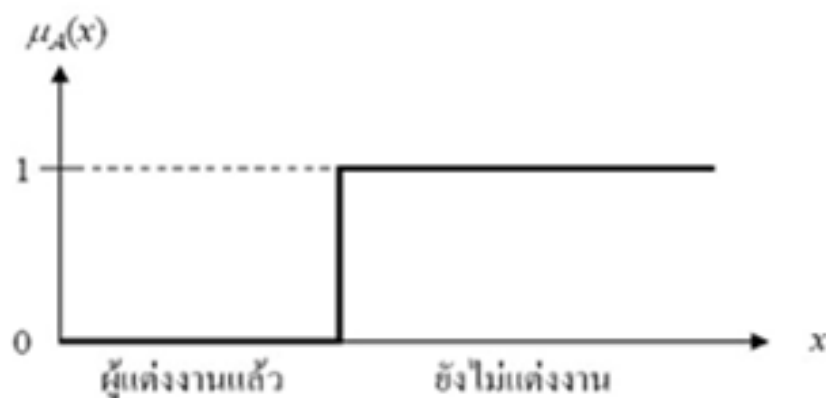
- กำหนดค่าความเป็นสมาชิกตามแนวคิดเลขฐานสอง
- เซตที่มีค่าความเป็นสมาชิกเป็น 0 หรือ 1 เท่านั้น
- ขอบเขตของเซตที่ตัดขาดจากกันแบบทันทีทันใด ไม่มีความต่อเนื่อง

$$\mu_A(x) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

เซตแบบต้นฉบับ (ต่อ)



ตัวอย่าง Crisp Set (Classical Set)



ตัวอย่างการแสดงความน่าจะเป็นสมาชิกของผู้ที่ยังไม่ได้แต่งงาน

Fuzzy Logic

“ทุกสิ่งบนโลกแห่งความเป็นจริง มิได้มีเฉพาะสิ่งที่มีความแน่นอนเท่านั้น แต่มีหลายสิ่ง หลายอย่าง หลายเหตุการณ์เกิดขึ้นอย่างไม่เที่ยงตรง อาจเป็นสิ่งที่คลุมเครือและไม่แน่นอน”

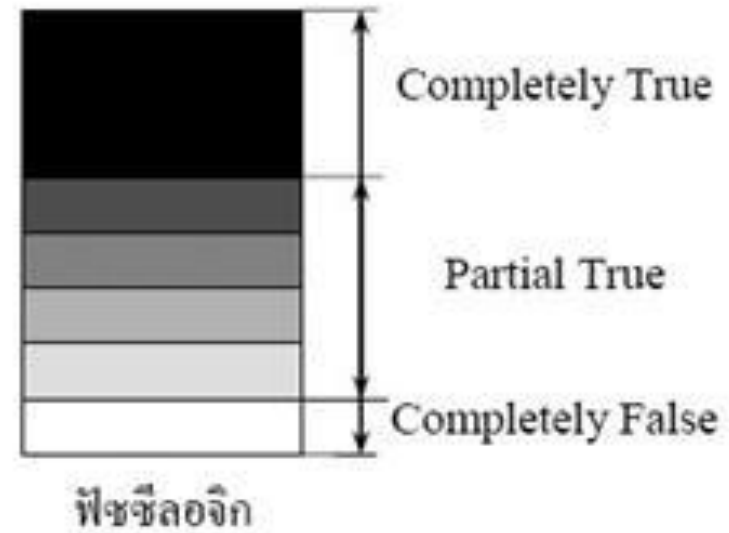
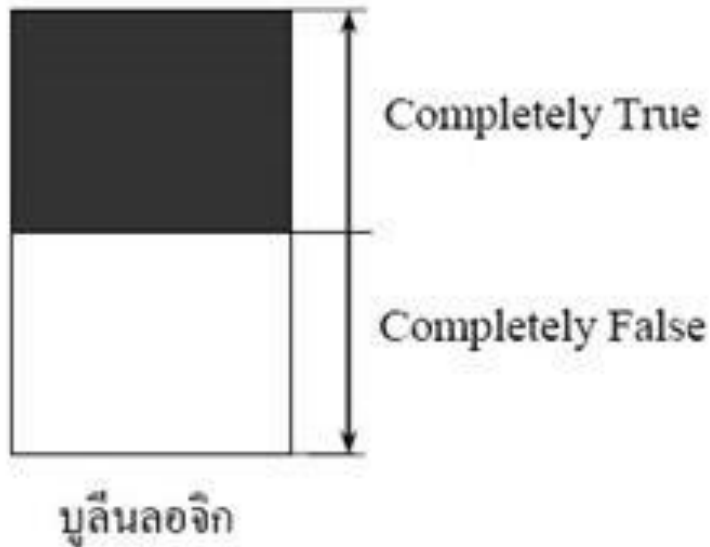
Fuzzy Set

Example

เซตของอายุคน อาจแบ่งเป็น **วัยทารก** **วัยเด็ก** **วัยรุ่น** **วัยกลางคน** และ **วัยชรา** แต่ละช่วงอายุคนไม่สามารถระบุได้แน่ชัดว่า **วัยทารก** กับ **วัยเด็ก** แยกจากกัน **แน่ชัดช่วงใด** **วัยทารก** อาจถูกตีความเป็นอายุระหว่าง 0 ถึง 1 ปี บางคนอาจตีความว่า **วัยทารก** อยู่ในช่วงอายุ 0 ถึง 2 ปี เซตของเหตุการณ์ที่ไม่แน่นอนเช่นนี้เรียกว่า “ฟัซซีเซต” (Fuzzy Set)

Fuzzy Logic

- ฟัซซี่ลอจิกมีลักษณะที่พิเศษกว่าตรรกะแบบจริงเท็จ (Boolean logic) โดยมีการเพิ่มแนวคิด **ความจริงบางส่วน (partial true)** เข้ามา ซึ่งจะมีความจริงอยู่ในช่วงระหว่างจริง (completely true) กับเท็จ (completely false)



Fuzzy set

- ถูกนำเสนอในรูปแบบของค่าขอบเขตที่คลุมเครือ
- ยอมให้มีค่าความเป็นสมาชิกของเซตมีได้มากกว่า 2 ค่า
 - ความเป็นสมาชิกจะมีค่าระหว่าง 0 – 1 **[0,1]**

โดย 0 หมายถึง ไม่มีความเป็นสมาชิกเลย

1 หมายถึง ความเป็นสมาชิกโดยสมบูรณ์

ระหว่าง 0 – 1 หมายถึง ความเป็นสมาชิกแค่บางส่วน

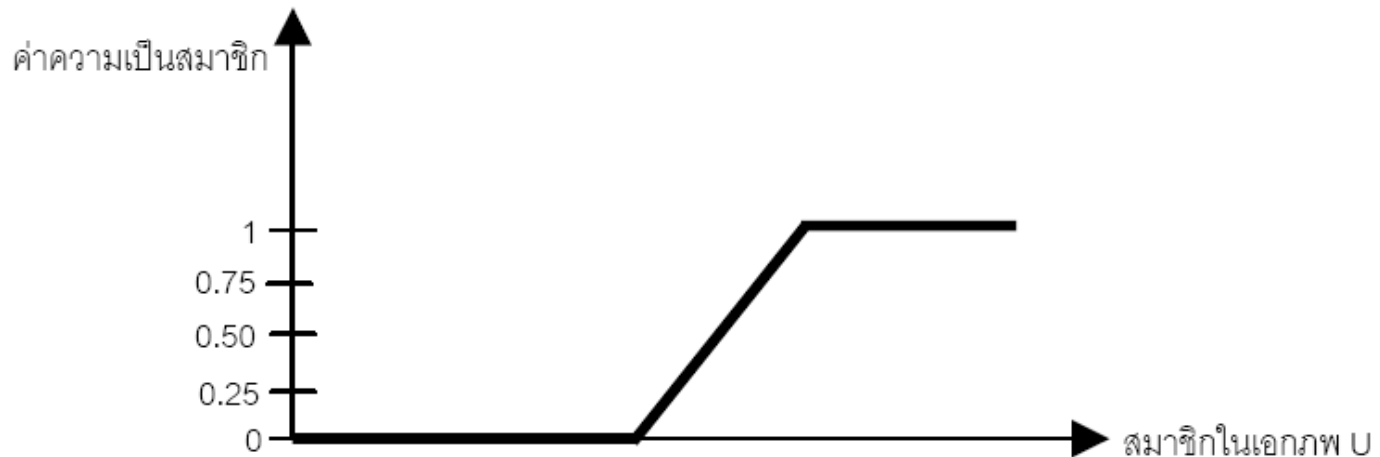
ดังนั้น ความเป็นสมาชิกของฟัซซีจะมีความต่อเนื่อง

Fuzzy set

ฟังก์ชันเซต A ของเอกภพ U ถูกกำหนดโดยฟังก์ชันความเป็นสมาชิก

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

กำหนดสมาชิกแต่ละตัว u ของเอกภพ คือ $\mu_A(u)$ โดยมีค่าอยู่ในช่วง $[0,1]$



กราฟแสดงความเป็นสมาชิก

Example

Query = “**cat** **and** **dog**”

$$d_1 = \{dog, cat, bird, zebra, zoo\}$$

$$d_2 = \{cat, kitty, fish\}$$

$$d_3 = \{dog, puppy, house, robber\}$$

$$d_4 = \{ant, sugar\}$$

Boolean = d_1

Fuzzy = d_1, d_2, d_3

Fuzzy set

ตัวดำเนินการหลัก ๆ ของฟัซซีเซต

Complement

$$\mu_{A^-}(x) = 1 - \mu_A(x)$$

Intersection

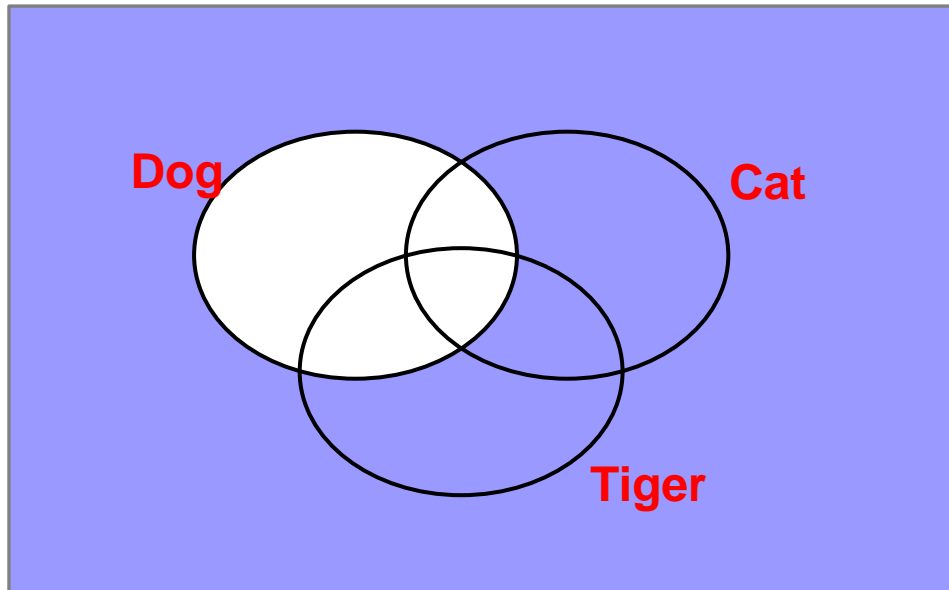
$$\mu_{A \cap B} = \mu_A(x) \wedge \mu_B(x) = \min(\mu_A(x), \mu_B(x))$$

Union

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

Fuzzy Logic

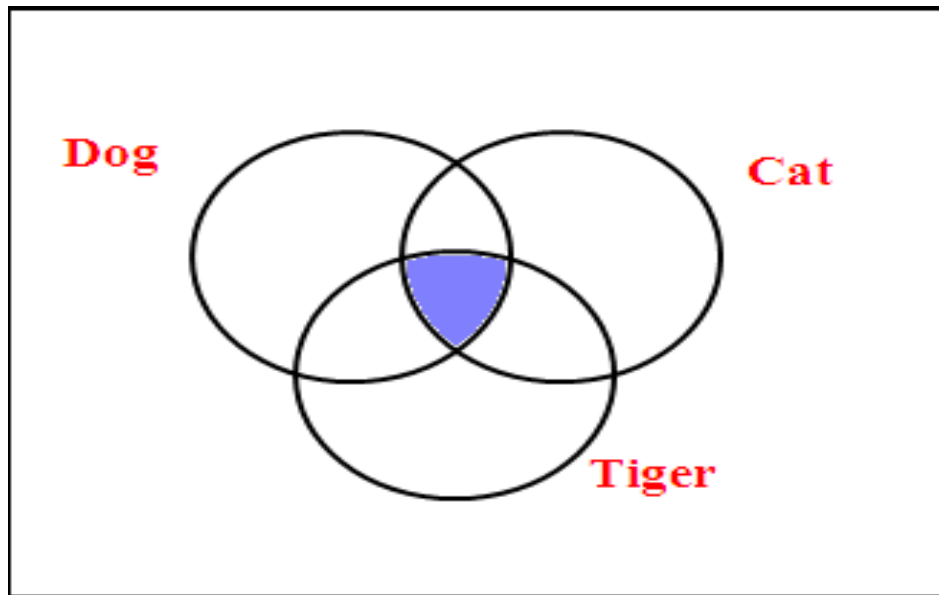
Complement



$$\mu_{\overline{Dog}}(x) = (1 - \mu_{Dog}(x))$$

Fuzzy Logic

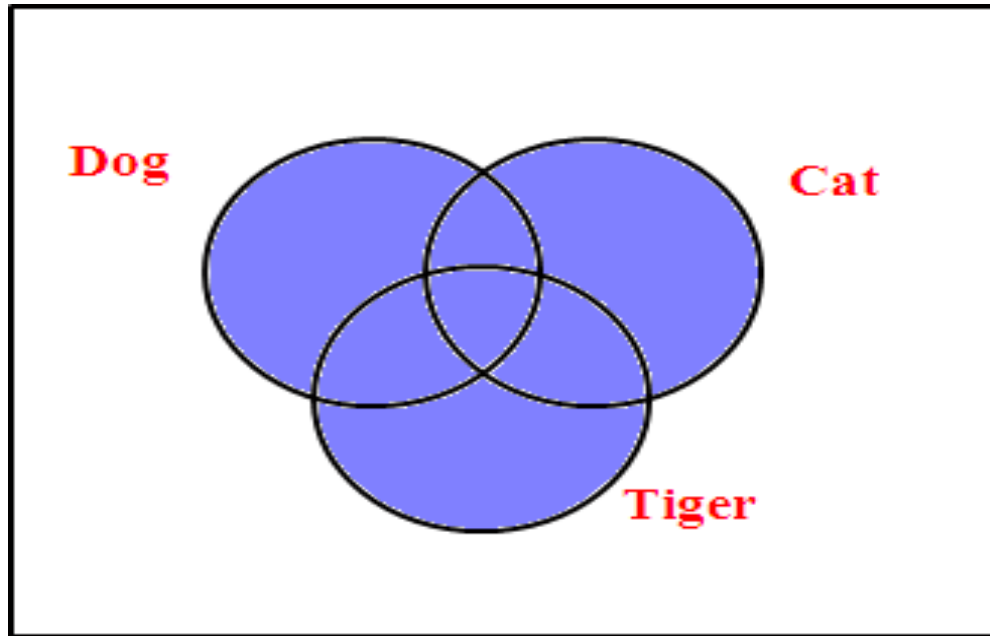
Intersection



$$\mu_{Dog \cap Cat \cap Tiger}(x) = \mu_{Dog}(x) \cdot \mu_{Cat}(x) \cdot \mu_{Tiger}(x)$$

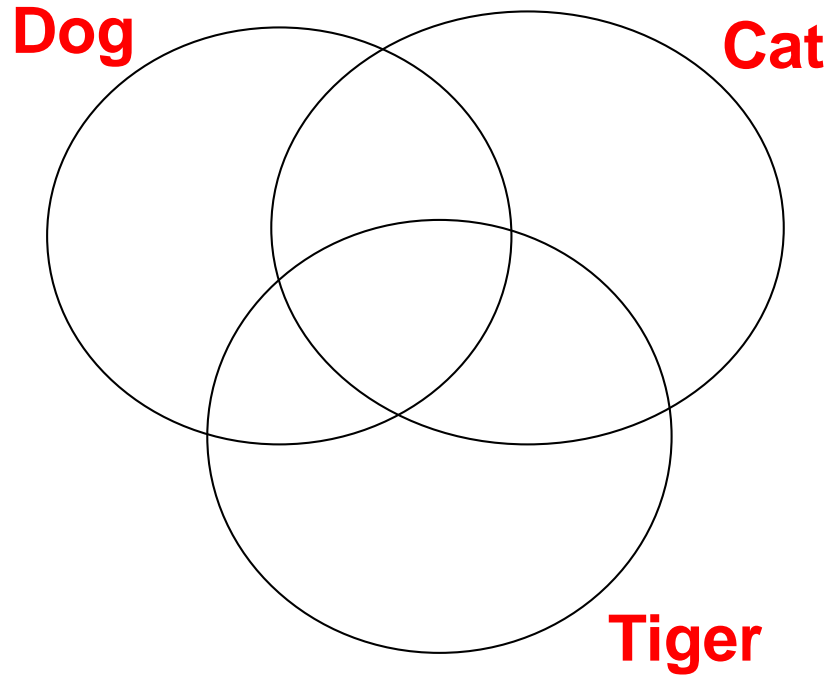
Fuzzy Logic

Union

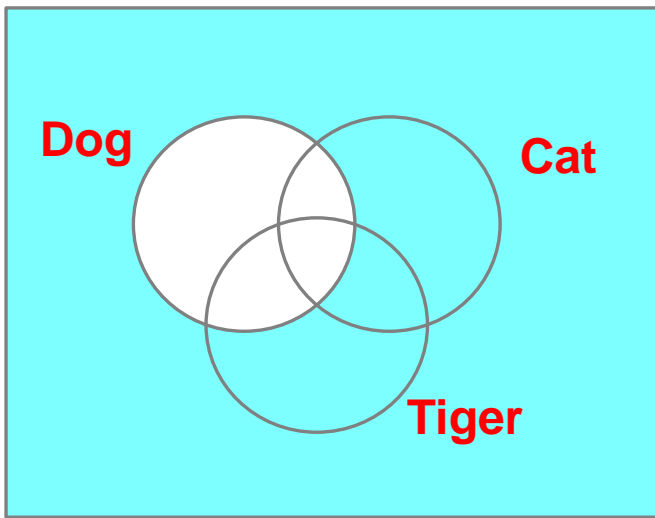


$$\mu_{Dog \cup Cat \cup Tiger}(x) = 1 - (1 - \mu_{Dog}(x)) \cdot (1 - \mu_{Cat}(x)) \cdot (1 - \mu_{Tiger}(x))$$

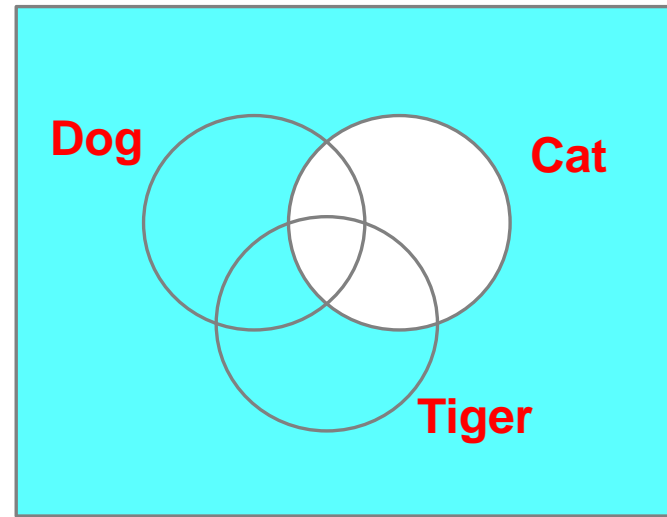
Example



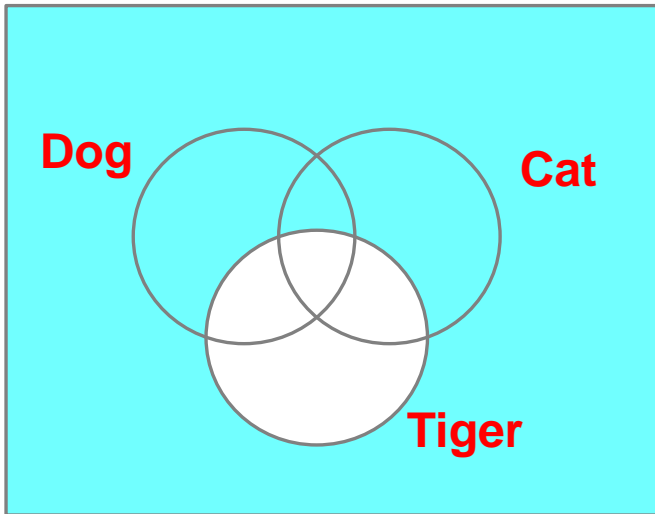
Dog OR Cat OR Tiger



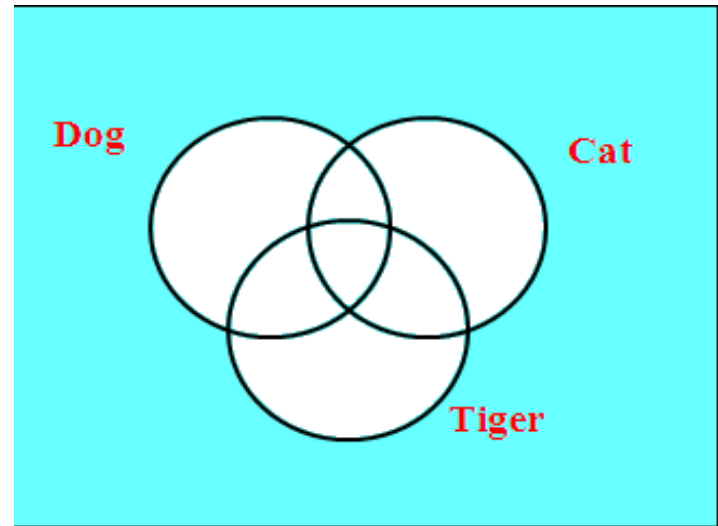
$$1 - \mu_{Dog}(x)$$



$$1 - \mu_{Cat}(x)$$

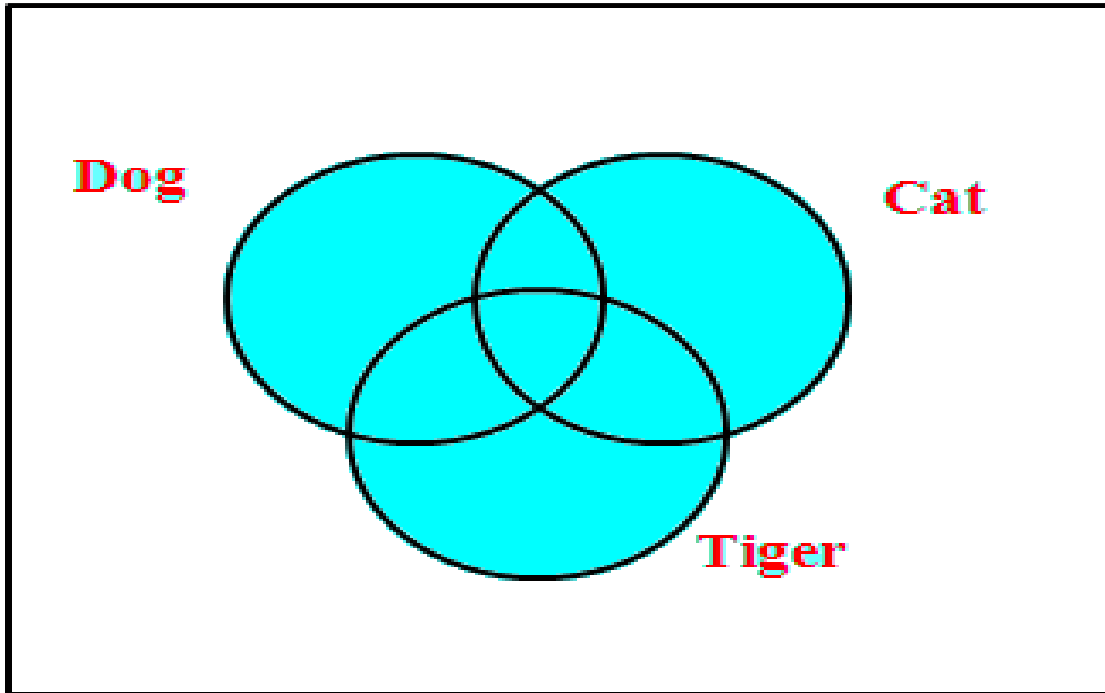


$$1 - \mu_{Tiger}(x)$$



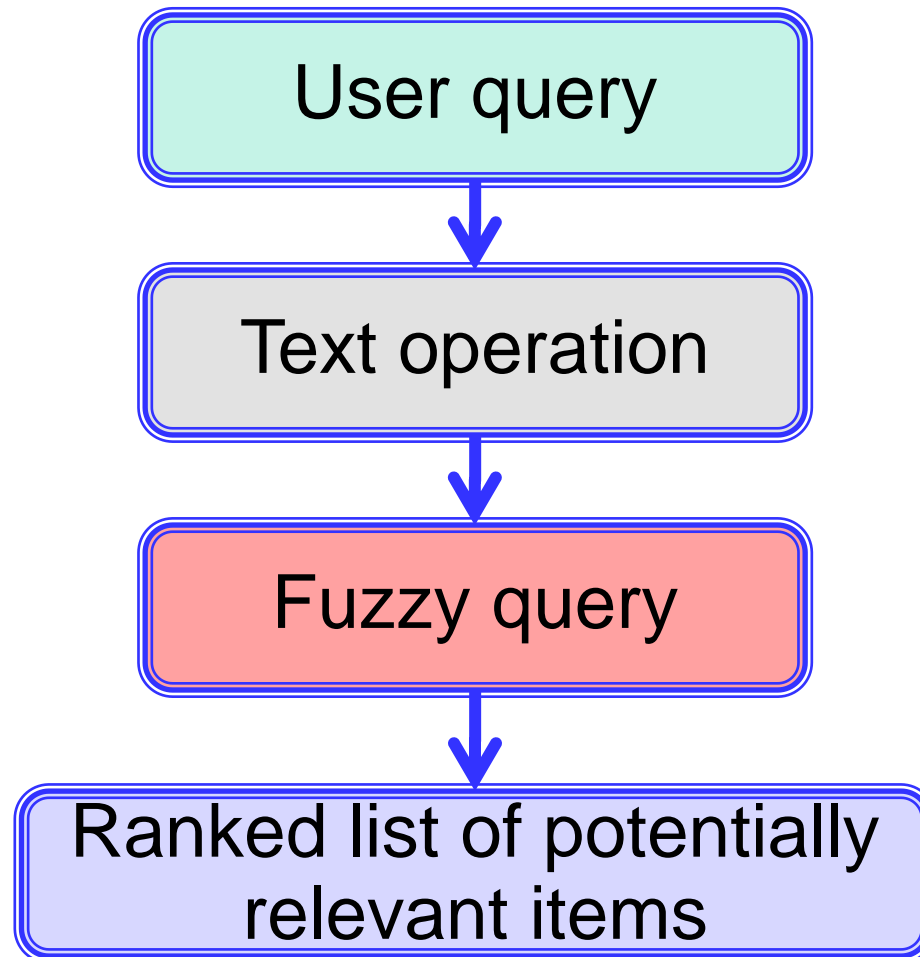
$$(1 - \mu_{Dog}(x)) \cdot (1 - \mu_{Cat}(x)) \cdot (1 - \mu_{Tiger}(x))$$

Example



$$\mu_{Dog \cup Cat \cup Tiger}(x) = 1 - (1 - \mu_{Dog}(x)) \cdot (1 - \mu_{Cat}(x)) \cdot (1 - \mu_{Tiger}(x))$$

Fuzzy Logic



Index Term Relationship

$$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

$c_{i,j}$ คือ ความสัมพันธ์ของคีย์เวิร์ด i กับ คีย์เวิร์ด j

$n_{i,j}$ คือ จำนวนเอกสารที่มีทั้งคีย์เวิร์ด i และคีย์เวิร์ด j

n_i คือ จำนวนเอกสารที่มีคีย์เวิร์ด i

n_j คือ จำนวนเอกสารที่มีคีย์เวิร์ด j

Example

Query = “I would like *cat* and have *dog*”

$$d_1 = \{dog, cat, bird, zebra, zoo\}$$

$$d_2 = \{cat, kitty, fish\}$$

$$d_3 = \{dog, puppy, house, robber\}$$

$$d_4 = \{ant, sugar\}$$

Example

Query = “*I would like **cat** and have **dog***”

I would like cat and have dog



I would like **cat** and have **dog**



Index term :

cat dog

Example

หาความสัมพันธ์ของคีย์เวิร์ดแต่ละตัว “cat” , ”dog”

$$c_{i,j} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}}$$

$c_{i,j}$ คือ ความสัมพันธ์ของคีย์เวิร์ด i กับ คีย์เวิร์ด j

$n_{i,j}$ คือ จำนวนเอกสารที่มีทั้งคีย์เวิร์ด i และคีย์เวิร์ด j

n_i คือ จำนวนเอกสารที่มีคีย์เวิร์ด i

n_j คือ จำนวนเอกสารที่มีคีย์เวิร์ด j

Example

cat → dog, bird, zebra, zoo, kitty, fish

$$c_{cat,dog} = \frac{n_{cat,dog}}{n_{cat} + n_{dog} - n_{cat,dog}} = \frac{1}{2 + 2 - 1} = \frac{1}{3} \approx 0.33$$

$$c_{cat,bird} = \frac{n_{cat,bird}}{n_{cat} + n_{bird} - n_{cat,bird}} = \frac{1}{2 + 1 - 1} = \frac{1}{2} = 0.5$$

$$c_{cat,zebra} = \frac{n_{cat,zebra}}{n_{cat} + n_{zebra} - n_{cat,zebra}} = \frac{1}{2 + 1 - 1} = \frac{1}{2} = 0.5$$

Example

$$c_{cat,zoo} = \frac{n_{cat,zoo}}{n_{cat} + n_{zoo} - n_{cat,zoo}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

$$c_{cat,kitty} = \frac{n_{cat,kitty}}{n_{cat} + n_{kitty} - n_{cat,kitty}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

$$c_{cat,fish} = \frac{n_{cat,fish}}{n_{cat} + n_{fish} - n_{cat,fish}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

Example

dog → cat, bird, zebra, zoo, puppy, house, robber

$$C_{dog,cat} = \frac{n_{dog,cat}}{n_{dog} + n_{cat} - n_{dog,cat}} = \frac{1}{2 + 2 - 1} = \frac{1}{3} \approx 0.33$$

$$C_{dog,bird} = \frac{n_{dog,bird}}{n_{dog} + n_{bird} - n_{dog,bird}} = \frac{1}{2 + 1 - 1} = \frac{1}{2} = 0.5$$

$$C_{dog,zebra} = \frac{n_{dog,zebra}}{n_{dog} + n_{zebra} - n_{dog,zebra}} = \frac{1}{2 + 1 - 1} = \frac{1}{2} = 0.5$$

Example

$$c_{dog,zoo} = \frac{n_{dog,zoo}}{n_{dog} + n_{zoo} - n_{dog,zoo}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

$$c_{dog,puppy} = \frac{n_{dog,puppy}}{n_{dog} + n_{puppy} - n_{dog,puppy}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

$$c_{dog,house} = \frac{n_{dog,house}}{n_{dog} + n_{house} - n_{dog,house}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

$$c_{dog,robber} = \frac{n_{dog,robber}}{n_{dog} + n_{robber} - n_{dog,robber}} = \frac{1}{2+1-1} = \frac{1}{2} = 0.5$$

Example

Index1	Index2	Relationship
Cat	Dog	0.33
Cat	Bird	0.5
Cat	Zebra	0.5
Cat	Zoo	0.5
Cat	Kitty	0.5
Cat	Fish	0.5

Example

Index1	Index2	Relationship
Dog	Cat	0.33
Dog	Bird	0.5
Dog	Zebra	0.5
Dog	Zoo	0.5
Dog	Puppy	0.5
Dog	House	0.5
Dog	Robber	0.5

Example

ระดับความเป็นสมาชิกของ index term กับแต่ละเอกสารในระบบ

$$cat = \{d_1 \rightarrow 1, d_2 \rightarrow 1, d_3 \rightarrow 0.33, d_4 \rightarrow 0\}$$

$$dog = \{d_1 \rightarrow 1, d_2 \rightarrow 0.33, d_3 \rightarrow 1, d_4 \rightarrow 0\}$$

$$d1 = \{dog, cat, bird, zebra, zoo\}$$

$$d2 = \{cat, kitty, fish\}$$

$$d3 = \{dog, puppy, house, robber\}$$

$$d4 = \{ant, sugar\}$$



Example

Union

$$\mu_{dog \cup cat}(x) = 1 - (1 - \mu_{dog}(x)) \cdot (1 - \mu_{cat}(x))$$

$$\mu_{dog \cup cat}(d_1) = 1 - (1 - \mu_{dog}(d_1)) \cdot (1 - \mu_{cat}(d_1)) = 1 - (1 - 1)(1 - 1) = 1$$

$$\mu_{dog \cup cat}(d_2) = 1 - (1 - \mu_{dog}(d_2)) \cdot (1 - \mu_{cat}(d_2)) = 1 - (1 - 0.33)(1 - 1) = 1$$

$$\mu_{dog \cup cat}(d_3) = 1 - (1 - \mu_{dog}(d_3)) \cdot (1 - \mu_{cat}(d_3)) = 1 - (1 - 1)(1 - 0.33) = 1$$

$$\mu_{dog \cup cat}(d_4) = 1 - (1 - \mu_{dog}(d_4)) \cdot (1 - \mu_{cat}(d_4)) = 1 - (1 - 0)(1 - 0) = 0$$

$$\mu_{dog \cup cat} = \{d_1 \rightarrow 1, d_2 \rightarrow 1, d_3 \rightarrow 1, d_4 \rightarrow 0\}$$

d_1, d_2, d_3, d_4

$d1 = \{dog, cat, bird, zebra, zoo\}$

$d2 = \{cat, kitty, fish\}$

$d3 = \{dog, puppy, house, robber\}$

$d4 = \{ant, sugar\}$

Example

Intersection

$$\mu_{dog \cap cat}(x) = \mu_{dog}(x) \cdot \mu_{cat}(x)$$

$$\mu_{dog \cap cat}(d_2) = \mu_{dog}(d_2) \cdot \mu_{cat}(d_2) = 0.33 \cdot 1 = 0.33$$

$$\mu_{dog \cap cat}(d_3) = \mu_{dog}(d_3) \cdot \mu_{cat}(d_3) = 1 \cdot 0.33 = 0.33$$

$$\mu_{dog \cap cat}(d_4) = \mu_{dog}(d_4) \cdot \mu_{cat}(d_4) = 0 \cdot 0 = 0$$

$$\mu_{dog \cap cat} = \{d_1 \rightarrow 1, d_2 \rightarrow 0.33, d_3 \rightarrow 0.33, d_4 \rightarrow 0\}$$

d_1, d_2, d_3, d_4

$d1 = \{dog, cat, bird, zebra, zoo\}$

$d2 = \{cat, kitty, fish\}$

$d3 = \{dog, puppy, house, robber\}$

$d4 = \{ant, sugar\}$

Example

Complement

$$\mu_{\overline{dog} \cap \overline{cat}}(x) = (1 - \mu_{dog}(x)) \cdot (1 - \mu_{cat}(x))$$

$$\mu_{\overline{dog} \cap \overline{cat}}(d_3) = (1 - \mu_{dog}(d_3)) \cdot (1 - \mu_{cat}(d_3)) = (1 - 1) \cdot (1 - 0.33) = 0$$

$$\mu_{\overline{dog} \cap \overline{cat}}(d_4) = (1 - \mu_{dog}(d_4)) \cdot (1 - \mu_{cat}(d_4)) = (1 - 0) \cdot (1 - 0) = 1$$

$$\mu_{\overline{dog} \cap \overline{cat}} = \{d_1 \rightarrow 0, d_2 \rightarrow 0, d_3 \rightarrow 0, d_4 \rightarrow 1\}$$

d_4, d_1, d_2, d_3

$d1 = \{dog, cat, bird, zebra, zoo\}$

$d2 = \{cat, kitty, fish\}$

$d3 = \{dog, puppy, house, robber\}$

$d4 = \{ant, sugar\}$

Example

- Union

$$d_1, d_2, d_3$$

- Intersection

$$d_1, d_2, d_3$$

- Complement

$$d_4$$

Extended Boolean Model

Boolean Model

เอกสาร	เนื้อหาของเอกสาร	ศัพท์ดรรชนีของเอกสาร
D ₁	สุนัขกินเหมือนกับที่แมวกิน	“สุนัข” “กิน” “แมว”
D ₂	สุนัขไม่ใช่หนู	“สุนัข” “หนู”
D ₃	หนูกินไม่มากนัก	“หนู” “กิน”
D ₄	แมวชอบเล่นกับงูและหนู	“แมว” “เล่น” “งู” “หนู”
D ₅	แมวชอบเล่น แต่ไม่กับแมวด้วยกัน	“แมว” “เล่น”

ถ้าต้องการค้น (แมว **AND** สุนัข) จะได้ผลการค้นเป็น **D₁** เท่านั้น

Boolean Model

- เนื่องจาก **Boolean Model** มีข้อเสียคือการไม่สนใจน้ำหนักของ **Keyword**
- **Vector Space Model** มีข้อเสียคือการเชื่อมต่อทางตรรกะทำได้ยาก

จึงได้มีความพยายามที่นำข้อดีของทั้งสองมารวมกัน ทำเป็น Model ใหม่ขึ้นมา เรียกว่า **Extended Boolean Model**



วิธีคำนวณค่าต่างๆ

น้ำหนักของ Keyword ในเอกสาร

น้ำหนักของ Keyword "i" ในเอกสาร "j"

$$w_{i,j} = tf_{normi,j} * idf_{normi}$$

normalized TF ของ Keyword "i" ในเอกสาร "j"

น้ำหนักของ Keyword ในเอกสาร

$$w_{i,j} = tf_{normi,j} * idf_{normi}$$



normalized IDF ของ Keyword “i” ในเอกสารทั้งหมด

น้ำหนักของ Keyword ในเอกสาร

$$w_{i,j} = tf_{normi,j} * idf_{normi}$$

$$tf_{normi,j} = \frac{tf_{i,j}}{tf_{\max i,j}}$$

$$idf_{normi} = \frac{idf_i}{idf_{\max g}}$$

น้ำหนักของ Keyword ในเอกสาร

tf คือจำนวนครั้งที่ Keyword นั้นปรากฏเอกสารที่สนใจ

tf ของ Keyword “i” ในเอกสาร “j”

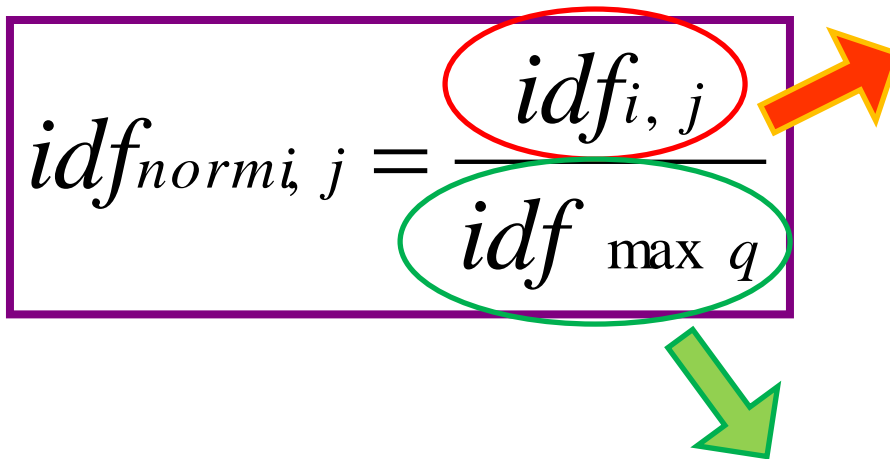
$$tf_{norm i, j} = \frac{tf_{i, j}}{tf_{\max i, j}}$$

ค่าสูงสุดที่หาได้ของ tf ของ Keyword ในเอกสาร “j”

น้ำหนักของ Keyword ในเอกสาร

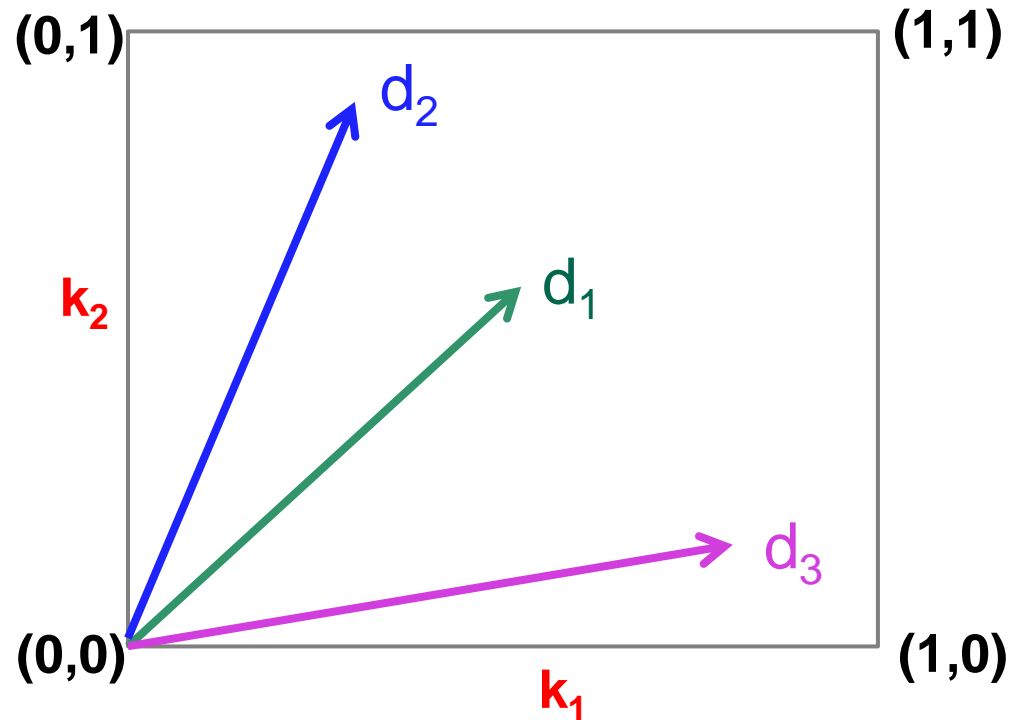
idf คือภาพรวมของการพบ Keyword ที่สนใจ โดยพิจารณาจากเอกสารทั้งหมดในระบบ

idf ของ Keyword “i” ในเอกสารทั้งหมด

$$idf_{norm i, j} = \frac{idf_{i, j}}{idf_{\max q}}$$


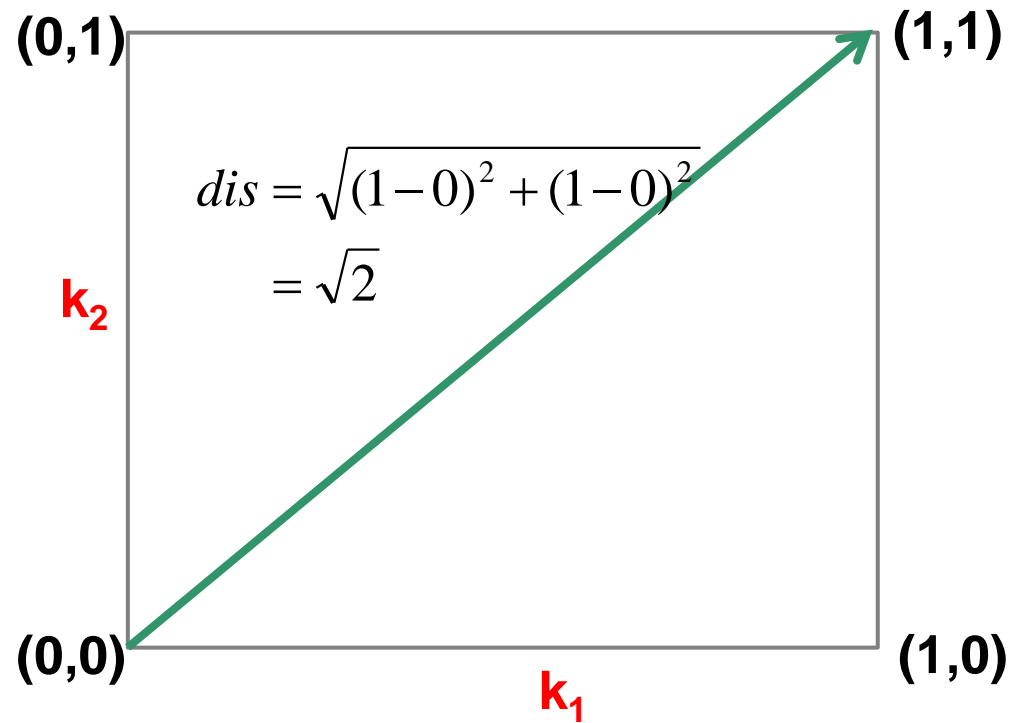
ค่าสูงสุดที่หาได้ของ idf ของ Keyword ในเอกสารทั้งหมด

Relevance



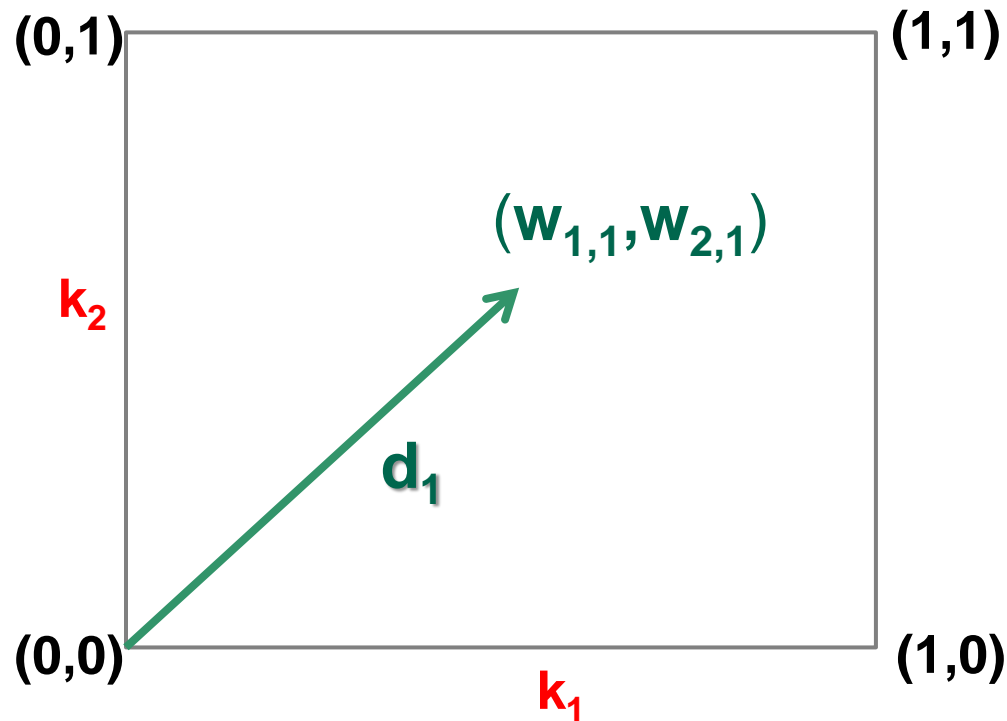
จุด $(0,0)$ คือจุดที่มีความตรงประเด็นน้อยที่สุด
จุด $(1,1)$ คือจุดที่มีความตรงประเด็นมากที่สุด

ระยะห่างสูงสุดของความตรงประเด็น (Relevance)



OR

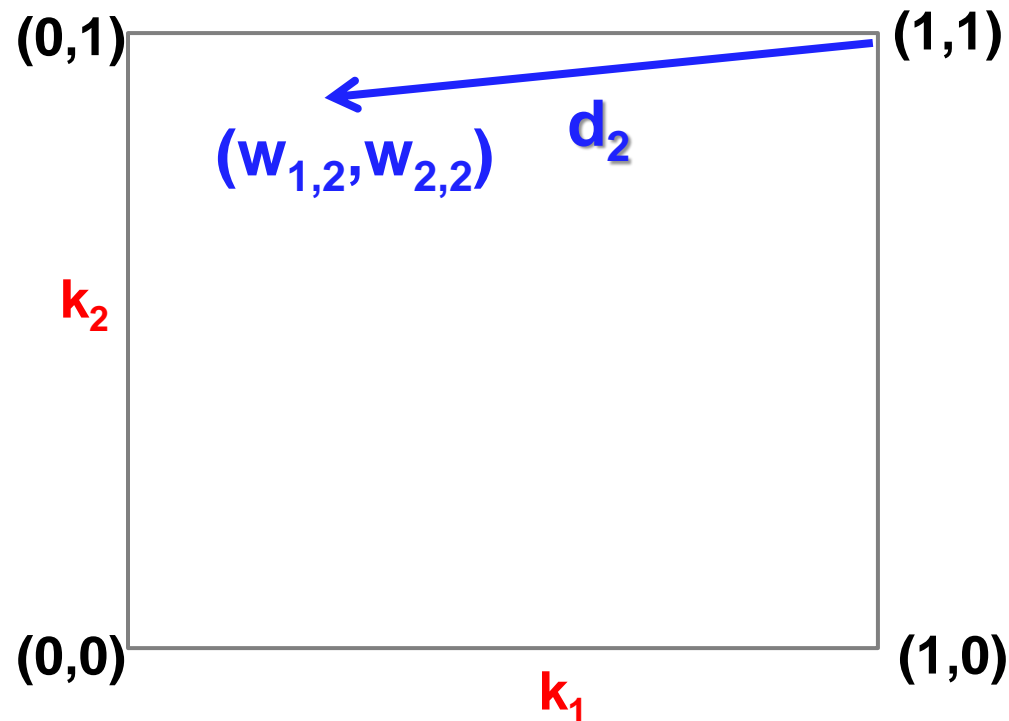
คำนวณระยะห่างจากจุด(0,0) ไปที่ (w_1, w_2) ของเอกสารที่สนใจ



$$dis_{or} = \sqrt{(W_{1,j}-0)^2 + (W_{2,j}-0)^2}$$

AND

คำนวณระยะห่างจากจุด(1,1) ไปที่ (w_1, w_2) ของเอกสารที่สนใจ



$$dis_{AND} = \sqrt{2} - \sqrt{(1-w_{1,j})^2 + (1-w_{2,j})^2}$$

ความตรงประเด็นของ Query

OR

$$dis_{or} = \sqrt{(W_{1,j})^2 + (W_{2,j})^2}$$

ระยะห่าง $\sqrt{2}$

ตรงประเด็น 1.00

ระยะห่าง dis

ตรงประเด็น $\frac{dis}{\sqrt{2}}$

$$sim(q_{or}, dj) = \sqrt{\frac{W_{1,j}^2 + W_{2,j}^2}{2}}$$

ความตรงประเด็นของ Query

AND

$$dis_{and} = \sqrt{2} - \sqrt{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}$$

ระยะห่าง $\sqrt{2}$ ตรงประเด็น 1.00
ระยะห่าง dis ตรงประเด็น $\frac{dis}{\sqrt{2}}$

$$sim(q_{and}, dj) = 1 - \sqrt{\frac{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}{2}}$$

Question

Q = Cat **OR** Not dog ???

$$\text{sim}(q_{or}, dj) = \sqrt{\frac{W_{cat,j}^2 + (1 - W_{dog,j})^2}{2}}$$

Q = (Dog **AND** Cat) **OR** Tiger ???

$$\text{sim}(q_{or}, dj) = \sqrt{\frac{W_{1,j}^2 + W_{tiger,j}^2}{2}}$$

$$\text{sim}(q_{or}, dj) = \sqrt{\frac{\left(1 - \sqrt{\frac{(1 - W_{dog,j})^2 + (1 - W_{cat,j})^2}{2}}\right)^2 + W_{tiger,j}^2}{2}}$$

EXAMPLE

ข้อ 2. สมมติในระบบมีเอกสาร 10 เอกสารดังนี้ (bird, cat, dog, tiger คือ Keyword ซึ่งไม่สัมพันธ์กัน)

- D1: {bird, cat, bird, cat, dog, dog, bird}
- D2: {cat, tiger, cat, dog}
- D3: {dog, bird, bird}
- D4: {cat, tiger}
- D5: {tiger, tiger, dog, tiger, cat}
- D6: {bird, cat, bird, cat, tiger, tiger, bird}
- D7: {bird, tiger, cat, dog}
- D8: {dog, cat, bird}
- D9: {cat, dog, tiger}
- D10: {tiger, tiger, tiger}

เด็กหญิงดาวิกาส่งคำเรียกค้น "รักแมวและสุนัข แต่ไม่รักเสือ" เข้าไปในระบบ จงตอบคำถาม

2.1 เพื่อให้ได้คำตอบในคำถาม 2.2 เด็กหญิงดาวิกาควรเลือกใช้โมเดลใดเพราะอะไร (เลือกได้เฉพาะตัวเลือกที่ให้มา)

- A) Probabilistic Model B) Fuzzy Model C) Extend Boolean Model D) Vector Model

2.2 ให้นักศึกษาแสดงวิธีคำนวณหา Ranking ของเอกสารทุกเอกสารในระบบ ตามที่เด็กหญิงดาวิกาต้องการ

(33 คะแนน) ข้อสอบ 1/2559

Answer

2.1 เลือกใช้ Extend Boolean Model เนื่องจากลักษณะของ Query เป็นแบบ Boolean และโจทย์กำหนดให้ Keyword ไม่สัมพันธ์กัน

ขั้นตอนที่ 1

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

D1: {bird, cat, bird, cat, dog, dog, bird}

D2: {cat, tiger, cat, dog}

D3: {dog, bird, bird}

D4: {cat, tiger}

D5: {tiger, tiger, dog, tiger, cat}

D6: {bird, cat, bird, cat, tiger, tiger, bird}

D7: {bird, tiger, cat, dog}

D8: {dog, cat, bird}

D9: {cat, dog, tiger}

D10: {tiger, tiger, tiger}

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger

Ranking

Doc8

Doc1

...

	Bird	Cat	Dog	Tiger	Max
Doc1	3	2	2	0	3
Doc2	0	2	1	1	2
Doc3	2	0	1	0	2
Doc4	0	1	0	1	1
Doc5	0	1	1	3	3
Doc6	3	2	0	2	3
Doc7	1	1	1	1	1
Doc8	1	1	1	0	1
Doc9	0	1	1	1	1
Doc10	0	0	0	3	3
n	5	8	7	7	

ขั้นตอนที่ 1

Only Doc1

$$tf_{bird} = \frac{3}{3} = 1.000$$

$$tf_{cat} = \frac{2}{3} = 0.667$$

$$tf_{dog} = \frac{2}{3} = 0.667$$

$$tf_{tiger} = \frac{0}{3} = 0.000$$

	Bird	Cat	Dog	Tiger	Max
Doc1	3	2	2	0	3
Doc2	0	2	1	1	2
Doc3	2	0	1	0	2
Doc4	0	1	0	1	1
Doc5	0	1	1	3	3
Doc6	3	2	0	2	3
Doc7	1	1	1	1	1
Doc8	1	1	1	0	1
Doc9	0	1	1	1	1
Doc10	0	0	0	3	3
n	5	8	7	7	

$$idf_{bird} = \log\left(\frac{10}{5}\right) = 0.301$$

$$idf_{cat} = \log\left(\frac{10}{8}\right) = 0.097$$

$$idf_{dog} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{tiger} = \log\left(\frac{10}{7}\right) = 0.155$$

$$idf_{norm, bird} = \frac{0.301}{0.301} = 1.000$$

$$idf_{norm, cat} = \frac{0.097}{0.301} = 0.322$$

$$idf_{norm, dog} = \frac{0.155}{0.301} = 0.515$$

$$idf_{norm, tiger} = \frac{0.155}{0.301} = 0.515$$

$$w_{bird} = 1.000 * 1.000 = 1.000$$

$$w_{cat} = 0.667 * 0.322 = 0.215$$

$$w_{dog} = 0.667 * 0.515 = 0.343$$

$$w_{tiger} = 0.000 * 0.515 = 0.000$$

น้ำหนักของแต่ละ **Keyword** ในแต่ละเอกสาร

	Bird	Cat	Dog	Tiger
Doc1	1.000	0.215	0.343	0.000
Doc2	0.000	0.322	0.257	0.257
Doc3	1.000	0.000	0.257	0.000
Doc4	0.000	0.322	0.000	0.515
Doc5	0.000	0.107	0.172	0.515
Doc6	1.000	0.215	0.000	0.343
Doc7	1.000	0.322	0.515	0.515
Doc8	1.000	0.322	0.515	0.000
Doc9	0.000	0.322	0.515	0.515
Doc10	0.000	0.000	0.000	0.515

ขั้นตอนที่ 2

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger

$$sim(q_{and}, dj) = 1 - \sqrt{\frac{(1 - W_{1,j})^2 + (1 - W_{2,j})^2}{2}}$$

	Bird	Cat	Dog	Tiger
Doc1	1.000	0.215	0.343	0.000
Doc2	0.000	0.322	0.257	0.257
Doc3	1.000	0.000	0.257	0.000
Doc4	0.000	0.322	0.000	0.515
Doc5	0.000	0.107	0.172	0.515
Doc6	1.000	0.215	0.000	0.343
Doc7	1.000	0.322	0.515	0.515
Doc8	1.000	0.322	0.515	0.000
Doc9	0.000	0.322	0.515	0.515
Doc10	0.000	0.000	0.000	0.515

$$sim(q_{and}, dj) = 1 - \sqrt{\frac{\left(1 - \left(1 - \sqrt{\frac{(1 - W_{Cat,j})^2 + (1 - W_{Dog,j})^2}{2}}\right)\right)^2 + (1 - (1 - W_{Tiger,j}))^2}{2}}$$

$$sim(q_{and}, d_1) = 1 - \sqrt{\frac{\left(1 - \left(1 - \sqrt{\frac{(1 - 0.215)^2 + (1 - 0.343)^2}{2}}\right)\right)^2 + (1 - (1 - 0.000))^2}{2}}$$

$$sim(q_{and}, d_1) = 0.488$$

ขั้นตอนที่ 3

Query = รักแมวและสุนัข แต่ไม่รักเสือ

Query = (Cat AND Dog) AND NOT Tiger

เอกสาร 10 เอกสารมีการแจกแจง Keyword ดังนี้

D1: {bird, cat, bird, cat, dog, dog, bird}

D2: {cat, tiger, cat, dog}

D3: {dog, bird, bird}

D4: {cat, tiger}

D5: {tiger, tiger, dog, tiger, cat}

D6: {bird, cat, bird, cat, tiger, tiger, bird}

D7: {bird, tiger, cat, dog}

D8: {dog, cat, bird}

D9: {cat, dog, tiger}

D10: {tiger, tiger, tiger}

	Sim
Doc1	0.488
Doc2	0.465
Doc3	0.377
Doc4	0.295
Doc5	0.291
Doc6	0.320
Doc7	0.447
Doc8	0.583
Doc9	0.447
Doc10	0.205

Ranking	Sim
Doc8	0.583
Doc1	0.488
Doc2	0.465
Doc7	0.447
Doc9	0.447
Doc3	0.377
Doc6	0.320
Doc4	0.295
Doc5	0.291
Doc10	0.205

Rank → Doc8, Doc1, Doc2, Doc7, Doc9, Doc3, Doc6, Doc4, Doc5, Doc10

Generalizes Vector Space Model (GVSM)

Basic Vector Space Model

Example

D1 =(2, 1, 0, 0)

D2 =(5, 1, 0, 0)

D3 =(1, 1, 1, 1)

D4 =(0, 0, 2, 2)

D5 =(0, 1, 1, 2)

D6 =(0, 0, 1, 1)

D7 =(0, 0, 1, 0)

D8 =(1, 1, 0, 0)

D9 =(2, 1, 1, 1)

D10=(0, 2, 2, 2)

D11=(1, 0, 2, 0)

D12=(0, 0, 2, 1)

$$q = 2k_1 + 3k_2 - k_3$$

$$W_{ij} = tf_{ij} * idfi = tf_{ij} * \log \left(\frac{N}{n_i} \right)$$

$$W_{iq} = \left(0.5 + \frac{0.5 * freqi_q}{Max(freqi_q)} \right) * \log \left(\frac{N}{n_i} \right)$$

$$sim(q, dj) = \frac{\sum_{i=1}^t (w_{ij} * wiq)}{\sqrt{\sum_{i=1}^t w_{ij}^2 * \sum_{i=1}^t w_{iq}^2}}$$

Generalizes Vector Space Model

D1 =(2, 1, 0, 0)

D2 =(5, 1, 0, 0)

D3 =(1, 1, 1, 1)

D4 =(0, 0, 2, 2)

D5 =(0, 1, 1, 2)

D6 =(0, 0, 1, 1)

D7 =(0, 0, 1, 0)

D8 =(1, 1, 0, 0)

D9 =(2, 1, 1, 1)

D10 =(0, 2, 2, 2)

D11 =(1, 0, 2, 0)

D12 =(0, 0, 2, 1)

$$q = 2k_1 + 3k_2 - k_3$$

D1 =(2, 1, 0, 0)

D2 =(5, 1, 0, 0)

D8 =(1, 1, 0, 0)

D3 =(1, 1, 1, 1)

D9 =(2, 1, 1, 1)

D4 =(0, 0, 2, 2)

D6 =(0, 0, 1, 1)

D12 =(0, 0, 2, 1)

D5 =(0, 1, 1, 2)

D10 =(0, 2, 2, 2)

D7 =(0, 0, 1, 0)

D11 =(1, 0, 2, 0)

ข้อเสียของ Vector Space Model

- Keyword บางส่วนอาจจะเป็นอิสระต่อกัน บางส่วนอาจจะเกี่ยวข้องกันและบางส่วนอาจจะเกี่ยวข้องกันมาก
- การอนุมานว่า Keyword เป็นอิสระจากกัน จึงเป็นการกระทำไม่ได้สอดคล้องกับความเป็นจริง
- Keyword ที่มีความหมายเดียวกัน จัดอยู่ในกลุ่มเดียวกัน

Generalizes Vector Space Model (GVSM)

- Keyword จะไม่ได้เป็นอิสระต่อกัน แต่จะเกี่ยวข้องกันในลักษณะใดลักษณะหนึ่ง โดยสังเกตจากการปรากฏร่วมกัน
- การปรากฏของ Keyword จะนำมาซึ่งการเปรียบเทียบความคล้ายหรือความต่างของเอกสารกับคำเรียกค้นที่เข้ามา
- GVSM ใช้หลักการเดียวกับ VSM ด้วยการคำนวณหาค่าความสอดคล้องของคำเรียกค้นกับเอกสารในระบบ แต่บน Vector Space ใหม่

GVSM Definition

- Definition Given the set $\{k_1, k_2, \dots, k_t\}$ of index terms in a collection, as before, let $w_{i,j}$ be the weight associated with the term-document pair $[k_i, d_j]$. If the $w_{i,j}$ weights are **all binary**, then all possible patterns of term co-occurrence (inside documents) can be represented by a set of **2^t minterms** given by

$$m_1 = (0, 0, \dots, 0),$$

$$m_2 = (1, 0, \dots, 0)$$

, ...,

$$m_{2^t} = (1, 1, \dots, 1)$$

Let $g_i(m_j)$ return the weight $\{0, 1\}$ of the index term k_i in the minterm m_j

GVSM Definition

- Definition Let us define the following set of vectors

- $\mathbf{m}_1 = (0, 0, \dots, 1)$
 $\mathbf{m}_2 = (0, 0, \dots, 1, 0)$
 \dots
 $\mathbf{m}_{2^{t-1}} = (1, 1, \dots, 1)$

where each vector \mathbf{m}_i is associated with the respective minterm \mathbf{m}_i .

GVSM Definition

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

$$k_i = \frac{\sum_{\forall r, g_i(m_r)} c_{i,r} m_r}{\sqrt{\sum_{\forall r, g_i(m_r)} c_{i,r}^2}}$$

$$k_i \bullet k_j = \sum_{\forall r | g_i(m_r)=1 \wedge g_j(m_r)=1} c_{i,r} \times c_{j,r}$$

GVSM Definition

$$d_j = \sum_i w_{i,j} k_i \quad \longrightarrow \quad d_j = \sum_r s_{j,r} m_r$$

$$q_j = \sum_i w_{i,q} k_i \quad \longrightarrow \quad q_j = \sum_r s_{q,r} m_r$$

$$sim(q, d_j) = \frac{\sum_{i=1}^t w_{i,j} \cdot w_{i,q}}{\sqrt{\sum_{i=1}^t w_{i,j}^2 \cdot \sum_{i=1}^t w_{i,q}^2}}$$



$$sim(q, d_j) = \frac{\sum_r s_{d,r} \cdot s_{q,r}}{\sqrt{\sum_r s_{d,r}^2 \cdot \sum_r s_{q,r}^2}}$$

Example

D1 =(2, 1, 0, 0) **m1**

D2 =(5, 1, 0, 0) **m1**

D3 =(1, 1, 1, 1) **m2**

D4 =(0, 0, 2, 2) **m3**

D5 =(0, 1, 1, 2) **m4**

D6 =(0, 0, 1, 1) **m3**

D7 =(0, 0, 1, 0) **m5**

D8 =(1, 1, 0, 0) **m1**

D9 =(2, 1, 1, 1) **m2**

D10=(0, 2, 2, 2) **m4**

D11=(1, 0, 2, 0) **m6**

D12=(0, 0, 2, 1) **m3**

$$q = 2k_1 + 3k_2 - k_3$$

m1=(1, 1, 0, 0)

m2=(1, 1, 1, 1)

m3=(0, 0, 1, 1)

m4=(0, 1, 1, 1)

m5=(0, 0, 1, 0)

m6=(1, 0, 1, 0)

Minterms: 6 minterms

All Weight → TF*IDF First (คำนวณจาก vector model)

$$\mathbf{m1}=(1, 1, 0, 0)$$

$$\mathbf{m2}=(1, 1, 1, 1)$$

$$\mathbf{m3}=(0, 0, 1, 1)$$

$$\mathbf{m4}=(0, 1, 1, 1)$$

$$\mathbf{m5}=(0, 0, 1, 0)$$

$$\mathbf{m6}=(1, 0, 1, 0)$$

$$\mathbf{D1}=(2, 1, 0, 0) \mathbf{m1}$$

$$\mathbf{D2}=(5, 1, 0, 0) \mathbf{m1}$$

$$\mathbf{D3}=(1, 1, 1, 1) \mathbf{m2}$$

$$\mathbf{D4}=(0, 0, 2, 2) \mathbf{m3}$$

$$\mathbf{D5}=(0, 1, 1, 2) \mathbf{m4}$$

$$\mathbf{D6}=(0, 0, 1, 1) \mathbf{m3}$$

$$\mathbf{D7}=(0, 0, 1, 0) \mathbf{m5}$$

$$\mathbf{D8}=(1, 1, 0, 0) \mathbf{m1}$$

$$\mathbf{D9}=(2, 1, 1, 1) \mathbf{m2}$$

$$\mathbf{D10}=(0, 2, 2, 2) \mathbf{m4}$$

$$\mathbf{D11}=(1, 0, 2, 0) \mathbf{m6}$$

$$\mathbf{D12}=(0, 0, 2, 1) \mathbf{m3}$$

$$k_i = \frac{\sum_{\forall r, g_i(m_r)} c_{i,r} m_r}{\sqrt{\sum_{\forall r, g_i(m_r)} c_{i,r}^2}}$$

$$k_1 = \frac{c_{1,1}m_1 + c_{1,2}m_2 + c_{1,6}m_6}{\sqrt{c_{1,1}^2 + c_{1,2}^2 + c_{1,6}^2}}$$

$$k_1 = \frac{8m_1 + 3m_2 + m_6}{\sqrt{64 + 9 + 1}} \\ = \frac{8m_1 + 3m_2 + m_6}{\sqrt{74}}$$

$$k_2 = \frac{c_{2,1}m_1 + c_{2,2}m_2 + c_{2,4}m_4}{\sqrt{c_{2,1}^2 + c_{2,2}^2 + c_{2,4}^2}}$$

$$k_2 = \frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}}$$

$$c_{i,r} = \sum_{d_j | g_l(d_j)=g_l(m_r), for.all.l} w_{i,j}$$

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

$$c_{1,2} = w_{1,3} + w_{1,9} = 1 + 2 = 3$$

$$c_{1,6} = w_{1,11} = 1$$

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

$$c_{2,2} = w_{2,3} + w_{2,9} = 1 + 1 = 2$$

$$c_{2,4} = w_{2,5} + w_{2,10} = 1 + 2 = 3$$

m1=(1, 1, 0, 0)
m2=(1, 1, 1, 1)
m3=(0, 0, 1, 1)
m4=(0, 1, 1, 1)
m5=(0, 0, 1, 0)
m6=(1, 0, 1, 0)

D1=(2, 1, 0, 0) m1
D2=(5, 1, 0, 0) m1
D3=(1, 1, 1, 1) m2
D4=(0, 0, 2, 2) m3
D5=(0, 1, 1, 2) m4
D6=(0, 0, 1, 1) m3
D7=(0, 0, 1, 0) m5
D8=(1, 1, 0, 0) m1
D9=(2, 1, 1, 1) m2
D10=(0, 2, 2, 2) m4
D11=(1, 0, 2, 0) m6
D12=(0, 0, 2, 1) m3

$$k_3 = \frac{c_{3,2}m_2 + c_{3,3}m_3 + c_{3,4}m_4 + c_{3,5}m_5 + c_{3,6}m_6}{\sqrt{c_{3,2}^2 + c_{3,3}^2 + c_{3,4}^2 + c_{3,5}^2 + c_{3,6}^2}}$$

$$c_{3,2} = w_{3,3} + w_{3,9} = 1 + 1 = 2$$

$$c_{3,3} = w_{3,4} + w_{3,6} + w_{3,12} = 2 + 1 + 2 = 5$$

$$c_{3,4} = w_{3,5} + w_{3,10} = 1 + 2 = 3$$

$$c_{3,5} = w_{3,7} = 1$$

$$c_{3,6} = w_{3,11} = 2$$

$$k_3 = \frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}}$$

$$k_4 = \frac{c_{4,2}m_2 + c_{4,3}m_3 + c_{4,4}m_4}{\sqrt{c_{4,2}^2 + c_{4,3}^2 + c_{4,4}^2}}$$

$$c_{4,2} = w_{4,3} + w_{4,9} = 1 + 1 = 2$$

$$c_{4,3} = w_{4,4} + w_{4,6} + w_{4,12} = 2 + 1 + 1 = 4$$

$$c_{4,4} = w_{4,5} + w_{4,10} = 2 + 2 = 4$$

$$k_4 = \frac{2m_2 + 4m_3 + 4m_4}{6}$$

$$D1=(2, 1, 0, 0) \text{ m1}$$

$$D2=(5, 1, 0, 0) \text{ m1}$$

$$D3=(1, 1, 1, 1) \text{ m2}$$

$$D4=(0, 0, 2, 2) \text{ m3}$$

$$D5=(0, 1, 1, 2) \text{ m4}$$

$$D6=(0, 0, 1, 1) \text{ m3}$$

$$D7=(0, 0, 1, 0) \text{ m5}$$

$$D8=(1, 1, 0, 0) \text{ m1}$$

$$D9=(2, 1, 1, 1) \text{ m2}$$

$$D10=(0, 2, 2, 2) \text{ m4}$$

$$D11=(1, 0, 2, 0) \text{ m6}$$

$$D12=(0, 0, 2, 1) \text{ m3}$$

$$q = 2k_1 + 3k_2 - k_3$$

$$q = 2 * \left(\frac{8m_1 + 3m_2 + m_6}{\sqrt{74}} \right) + 3 * \left(\frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}} \right) - \left(\frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}} \right)$$

$$q = 3.779m_1 + 1.672m_2 - 0.762m_3 + 1.461m_4 - 0.152m_5 - 0.073m_6$$

$$d_1 = 2k_1 + k_2$$

$$d_1 = 2 * \left(\frac{8m_1 + 3m_2 + m_6}{\sqrt{74}} \right) + \left(\frac{3m_1 + 2m_2 + 3m_4}{\sqrt{22}} \right)$$

$$d_1 = 2.50m_1 + 1.124m_2 + 0.640m_4 + 0.232m_6$$

$$sim(q, d_j) = \frac{\sum_r S_{d,r} \cdot S_{q,r}}{\sqrt{\sum_r S_{d,r}^2 \cdot \sum_r S_{q,r}^2}}$$

$$sim(q, d_1) = \frac{2.50 * 3.779 + 1.124 * 1.672 + 0.640 * 1.461 - 0.232 * 0.073}{\sqrt{(2.50^2 + 1.124^2 + 0.640^2 + 0.232^2) * (3.779^2 + 1.672^2 + 0.762^2 + 1.461^2 + 0.152^2 + 0.073^2)}} \\ = 0.974$$

$$D1=(2, 1, 0, 0) \text{ m1}$$

$$D2=(5, 1, 0, 0) \text{ m1}$$

$$D3=(1, 1, 1, 1) \text{ m2}$$

$$D4=(0, 0, 2, 2) \text{ m3}$$

$$D5=(0, 1, 1, 2) \text{ m4}$$

$$D6=(0, 0, 1, 1) \text{ m3}$$

$$D7=(0, 0, 1, 0) \text{ m5}$$

$$D8=(1, 1, 0, 0) \text{ m1}$$

$$D9=(2, 1, 1, 1) \text{ m2}$$

$$D10=(0, 2, 2, 2) \text{ m4}$$

$$D11=(1, 0, 2, 0) \text{ m6}$$

$$D12=(0, 0, 2, 1) \text{ m3}$$

$$q = 3.779m_1 + 1.672m_2 - 0.762m_3 + 1.461m_4 - 0.152m_5 - 0.073m_6$$

$$d_4 = 2k_3 + 2k_4$$

$$d_4 = 2 * \left(\frac{2m_2 + 5m_3 + 3m_4 + m_5 + 2m_6}{\sqrt{43}} \right) + 2 * \left(\frac{2m_2 + 4m_3 + 4m_4}{6} \right)$$

$$d_4 = 1.277m_2 + 2.858m_3 + 2.248m_4 + 0.305m_5 + 0.610m_6$$

$$sim(q, d_j) = \frac{\sum_r S_{d,r} \cdot S_{q,r}}{\sqrt{\sum_r S_{d,r}^2 \cdot \sum_r S_{q,r}^2}}$$

$$sim(q, d_4) = \frac{1.277 * 1.672 - 2.858 * 0.762 + 2.248 * 1.461 - 0.305 * 0.152 - 0.610 * 0.073}{\sqrt{(1.277^2 + 2.858^2 + 2.248^2 + 0.305^2 + 0.610^2) * (3.779^2 + 1.672^2 + 0.762^2 + 1.461^2 + 0.152^2 + 0.073^2)}} \\ = 0.181$$

Degree of similarity

Rank

Doc	Sim
D1	0.974
D2	0.952
D3	0.697
D4	0.181
D5	0.419
D6	0.181
D7	0.124
D8	0.981
D9	0.806
D10	0.485
D11	0.494
D12	0.162

D8	
D1	m1
D2	
D9	
D3	m2
D11	m6
D10	
D5	m4
D4	
D6	m3
D12	
D7	m5

D1 =(2, 1, 0, 0) m1
 D2 =(5, 1, 0, 0) m1
 D3 =(1, 1, 1, 1) m2
 D4 =(0, 0, 2, 2) m3
 D5 =(0, 1, 1, 2) m4
 D6 =(0, 0, 1, 1) m3
 D7 =(0, 0, 1, 0) m5
 D8 =(1, 1, 0, 0) m1
 D9 =(2, 1, 1, 1) m2
 D10=(0, 2, 2, 2) m4
 D11=(1, 0, 2, 0) m6
 D12=(0, 0, 2, 1) m3

$$q = 2k_1 + 3k_2 - k_3$$

Generalizes Vector Space Model (GVSM)

Conclusions

- เอกสารที่มี minterm เดียวกันจะมีความตรงประเด็นใกล้เคียงกัน เนื่องจากลักษณะการปรากฏของ Keyword มีความคล้ายกัน
- Keyword อาจมีความเกี่ยวข้องกันได้ เช่น Cat และ Tiger (มีโอกาas เป็นไปได้สูงที่เอกสารที่มี Cat อาจมี Tiger อยู่ด้วย) ซึ่งหากมีการเรียกค้น Keyword ใด Keyword หนึ่ง เอกสารที่มีอีก Keyword หนึ่งก็จะตรงประเด็นด้วย