Mathematical Foundations

464.561A Models and Technologies for Information Services

Jonghun Park

jonghun@snu.ac.kr

Dept. of Industrial Eng.

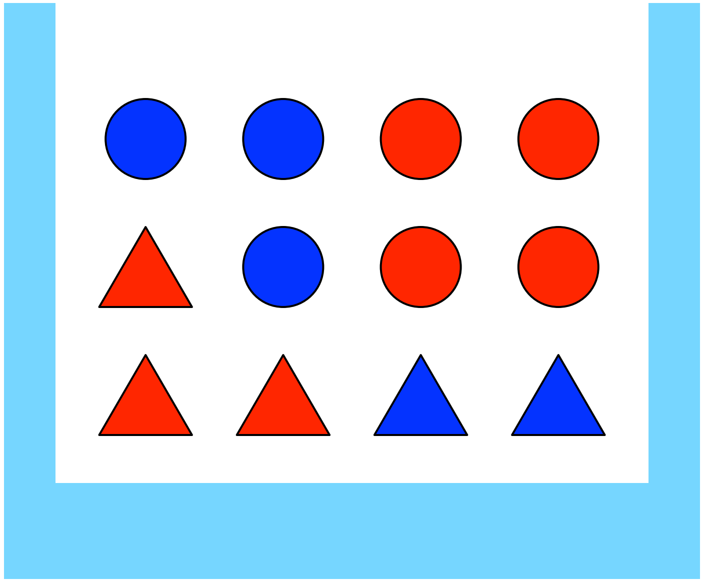
Seoul National University

3/7/11

Table of Contents

Elementary Probability Theory

Probability



Discrete Probability

distribution function p for random variable X:

$$p(x) \geq 0, \forall x \in \mathcal{X}$$

$$\sum_{x \in \mathcal{X}} p(x) = 1$$

probability of $E \subseteq \mathcal{X}$

$$P(E) = \sum_{x \in E} p(x)$$

Properties

$$P(E) \geq 0, \forall E \subset \mathcal{X}$$

$$P(\mathcal{X}) = 1$$

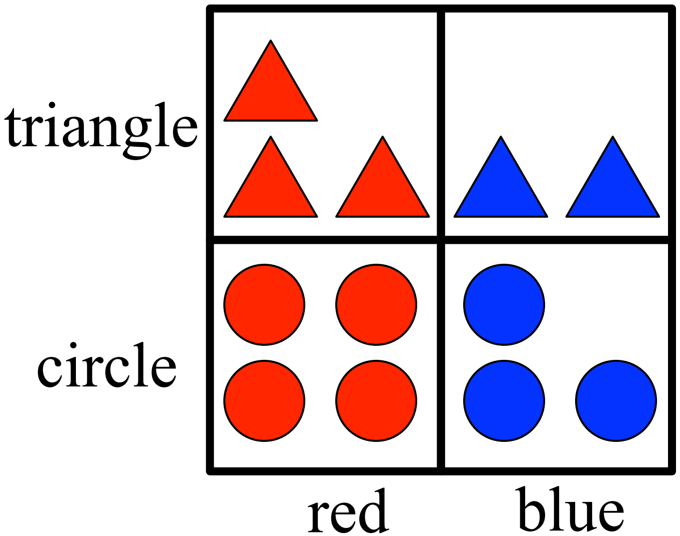
$$E \subset F \subset \mathcal{X} \Rightarrow P(E) \leq P(F)$$

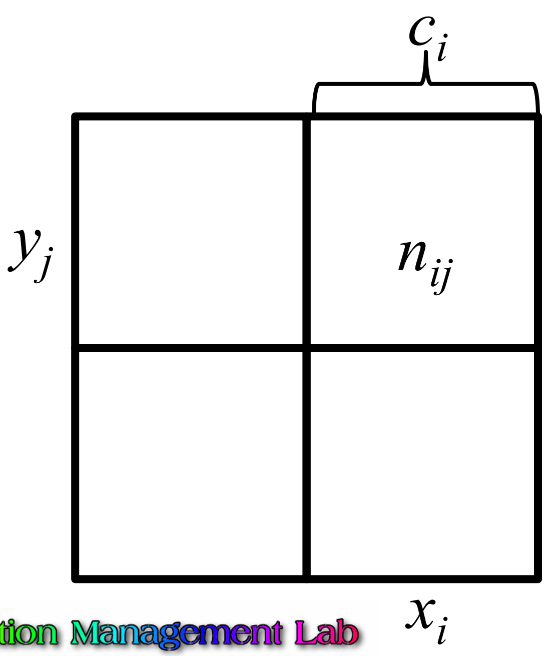
A and B are disjoint subsets of $\mathcal{X}$

$$\Rightarrow P(A \cup B) = P(A) + P(B)$$

$$P(\bar{A}) = 1 - P(A)$$

Marginal, Joint, & Conditional Prob.





Marginal Probability

$$P(X = x_i) = \frac{c_i}{N}$$

Joint Probability

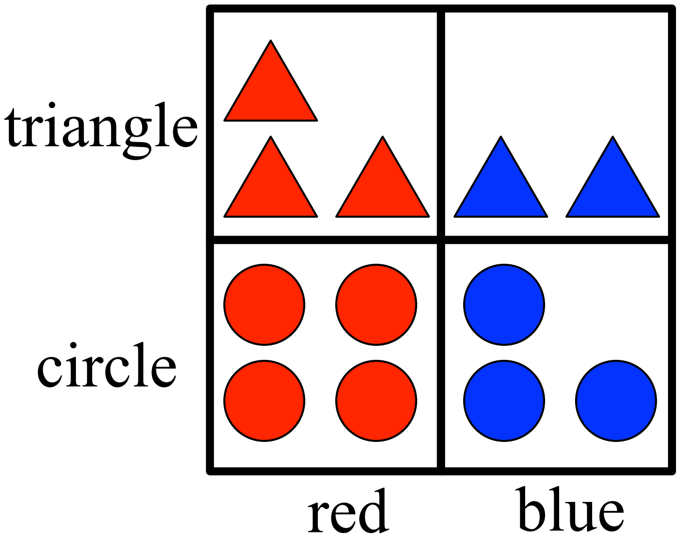
$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

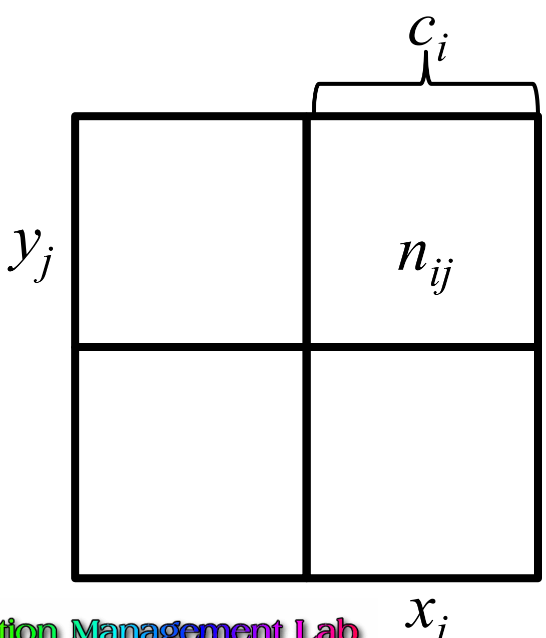
Conditional Probability

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

(Note) $\sum_Y P(Y|X) = 1$

Sum Rule & Product Rule





Sum Rule

$$P(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} P(X = x_i, Y = y_j)$$

$$\boxed{P(X) = \sum_{Y} P(X, Y)}$$

Product Rule

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= P(Y = y_j | X = x_i) P(X = x_i)$$

$$\boxed{P(X, Y) = P(Y|X)P(X)}$$

Bayes' Rule

$$P(X,Y) = P(Y,X)$$

$$P(Y|X)P(X) = P(X|Y)P(Y) \qquad \text{(product rule)}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

$$= \frac{P(X|Y)P(Y)}{\sum_Y P(X,Y)} \qquad \text{(sum rule)}$$

$$\boxed{= \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)}} \qquad \text{(product rule)}$$

Chain Rule

$$P(X, Y|Z) = \frac{P(X, Y, Z)}{P(Z)}$$

$$= \frac{P(X, Y, Z)}{P(Y, Z)} \cdot \frac{P(Y, Z)}{P(Z)}$$

$$= P(X|Y, Z) \cdot P(Y|Z)$$

$$\boxed{P(X, Y|Z) = P(X|Y, Z) \cdot P(Y|Z)}$$

Chain Rule: Generalization

$$P(X_1, ..., X_n)$$

$$= P(X_n|X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1})$$

$$= P(X_n|X_1, \ldots, X_{n-1})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2})$$

$$= P(X_n|X_1, \ldots, X_{n-1}) \cdots P(X_3|X_1, X_2)P(X_2|X_1)P(X_1)$$

Independence & Conditional Independence

$$P(X,Y) = P(X) \cdot P(Y), \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$$

$$\therefore P(X|Y) = P(X)$$

$$P(X,Y|Z) = P(X|Z) \cdot P(Y|Z), \forall (x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$$

denoted as $X \perp Y|Z$

Expectation

$$E(X) = \sum_{x \in \mathcal{X}} x p(x)$$

$$E(\phi(X)) = \sum_{x \in \mathcal{X}} \phi(x) p(x)$$

where $\phi : \mathcal{X} \rightarrow \Re$

Conditional Expectation

$$E(X|Y = y_j) \triangleq \sum_i x_i p(x_i|y_j)$$

$$E(X) = \sum_j E(X|Y = y_j)p(y_j)$$

$$E(\phi(X)|Y = y_j) = \sum_i \phi(x_i)p(x_i|y_j)$$

Variance

$$V(X) \triangleq E((X - E(X)^2)$$

$$= \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

$$\text{where } \mu = E(X)$$

MAP (Maximum A Posteriori)

estimate $\widehat{X}_{MAP}$ of X, given (observed) $Y = y_j$:

$$\widehat{X}_{MAP} \triangleq \arg\max_{x \in \mathcal{X}} p(x|y_j)$$

ML (Maximum Likelihood)

estimate $\widehat{X}_{ML}$ of X, given (observed) $Y = y_j$:

$$\widehat{X}_{ML} \triangleq \arg\max_{x \in \mathcal{X}} p(y_j | x)$$

Essential Information Theory

Entropy

measures the amount of information in a random variable

$$p(x) \triangleq P(X = x), \forall x \in \mathcal{X}$$

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

$$= E\left(\log \frac{1}{p(X)}\right)$$

Joint Entropy & Conditional Entropy

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$$

$$\begin{aligned}
H(Y|X) &\triangleq \sum_{x\in\mathcal{X}} p(x)H(Y|X=x) \\
&= \sum_{x\in\mathcal{X}} p(x)\left[-\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x)\right] \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)
\end{aligned}$$

Chain Rule for Entropy

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(p(x)p(y|x)\right) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(\log p(x)+\log p(y|x)\right) \\
&= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) \\
&= H(X) + H(Y|X)
\end{aligned}
$$

Chain Rule for Entropy (General Case)

$$H(X_1, \ldots, X_n)$$

$$= H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_{n-1})$$

Mutual Information (MI)

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

$$MI(X,Y) \triangleq H(X) - H(X|Y) = H(Y) - H(Y|X)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

note:

$$MI(X,X) = H(X) - H(X|X) = H(X)$$

Pointwise MI

$$
\begin{aligned}
MI(x,y) &= \log \frac{p(x,y)}{p(x)p(y)} \\
&= \log \frac{p(x|y)}{p(x)} \\
&= \log \frac{p(y|x)}{p(y)}
\end{aligned}
$$

Conditional MI

$$MI(X, Y|Z) \triangleq H(X|Z) - H(X|Y, Z)$$

Kullback-Leibler (KL) Divergence

$p(x)$, $q(x)$: probability mass functions

$$D(p||q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

$$= E_p \left(\log \frac{p(X)}{q(X)} \right)$$

where

$$0 \log \frac{0}{q} = 0$$

MI and KL Divergence

$$MI(X,Y) = D(p(x,y)||p(x)p(y))$$

Conditional KL Divergence

$$D(p(y|x)||q(y|x)) = \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log \frac{p(y|x)}{q(y|x)}$$

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Event Detection

Chi-Square Test

	Class 1	Class 2	Totals
Population 1	n_{11}	n_{12}	n_{1*}
Population 2	n_{21}	n_{22}	n_{2*}
Totals	n_{*1}	n_{*2}	$N = n_{1*} + n_{2*}$

$$\chi^2 = \frac{N \times (n_{11} \times n_{22} - n_{12} \times n_{21})^2}{n_{1*} \times n_{2*} \times n_{*1} \times n_{*2}}$$

Reject H_0: $p_1 = p_2$ when $\chi^2 \geq \theta$

Chi-Square Test Example

	# of People Who Changed a Character Name	# of People Who Didn't Change a Character Name	Totals
Before AionTem	13	73	86
After AionTem	17	57	74
Totals	30	130	160

$$\chi^2 = \frac{160 \times (13 \times 57 - 73 \times 17)^2}{86 \times 74 \times 30 \times 130} \approx 1.61 (< 3.84)$$

"no effect" at significance level = 0.05

Chi-Square Test for Hot Topic Detection

	# of docs containing term t_i	# of docs not containing term t_i	Totals
current time slot k	n_{11}	n_{12}	n_{1*}
previous H time slots	n_{21}	n_{22}	n_{2*}
Totals	n_{*1}	n_{*2}	n_{**}

$$n_{11} = df_i(k) \qquad n_{12} = N(k) - df_i(k)$$

$$n_{21} = \sum_{l=k-H}^{k-1} df_i(l) \quad n_{22} = \sum_{l=k-H}^{k-1} N(l) - \sum_{l=k-H}^{k-1} df_i(l)$$

$$\chi^2 = \frac{n** \times (|n_{11} \times n_{22} - n_{12} \times n_{21}| - \frac{1}{2} \times Y \times n_{**})^2}{n_{1*} \times n_{2*} \times n_{*1} \times n_{*2}}$$

MI (Mutual Information)

	# of docs containing term t_i	**# of docs not containing term t_i**	Totals
current time slot *k*	n_{11}	n_{12}	n_{1*}
previous *H* time slots	n_{21}	n_{22}	n_{2*}
Totals	n_{*1}	n_{*2}	n_{**}

$$MI(X,Y) = \log \frac{P(X,Y)}{P(X)P(Y)}$$

$$MI(t_i, s_k) = \log \frac{P(t_i \wedge s_k)}{P(t_i)P(s_k)} = \frac{\frac{n_{11}}{n_{**}}}{\frac{n_{11}+n_{21}}{n_{**}} \cdot \frac{n_{11}+n_{12}}{n_{**}}}$$

$$= \log \frac{n_{11} \cdot n_{**}}{(n_{11}+n_{12}) \cdot (n_{11}+n_{21})}$$

KL Divergence

	# of docs containing term t_i	# of docs not containing term t_i	Totals
current time slot *k*	n_{11}	n_{12}	n_{1*}
previous *H* time slots	n_{21}	n_{22}	n_{2*}
Totals	n_{*1}	n_{*2}	n_{**}

$$D(P||Q) = E_P\left(\log \frac{P(X)}{Q(X)}\right) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

$$MI(X,Y) = D(P(x,y)||P(x)P(y))$$

$$\sum_{i=1,2} \sum_{j=1,2} \frac{n_{ij}}{n_{**}} \log \frac{\frac{n_{ij}}{n_{**}}}{\frac{n_{*j}}{n_{**}} \cdot \frac{n_{i*}}{n_{**}}}$$

Similarity

Similarity

$$\sigma : O \times O \to \Re$$

s.t.

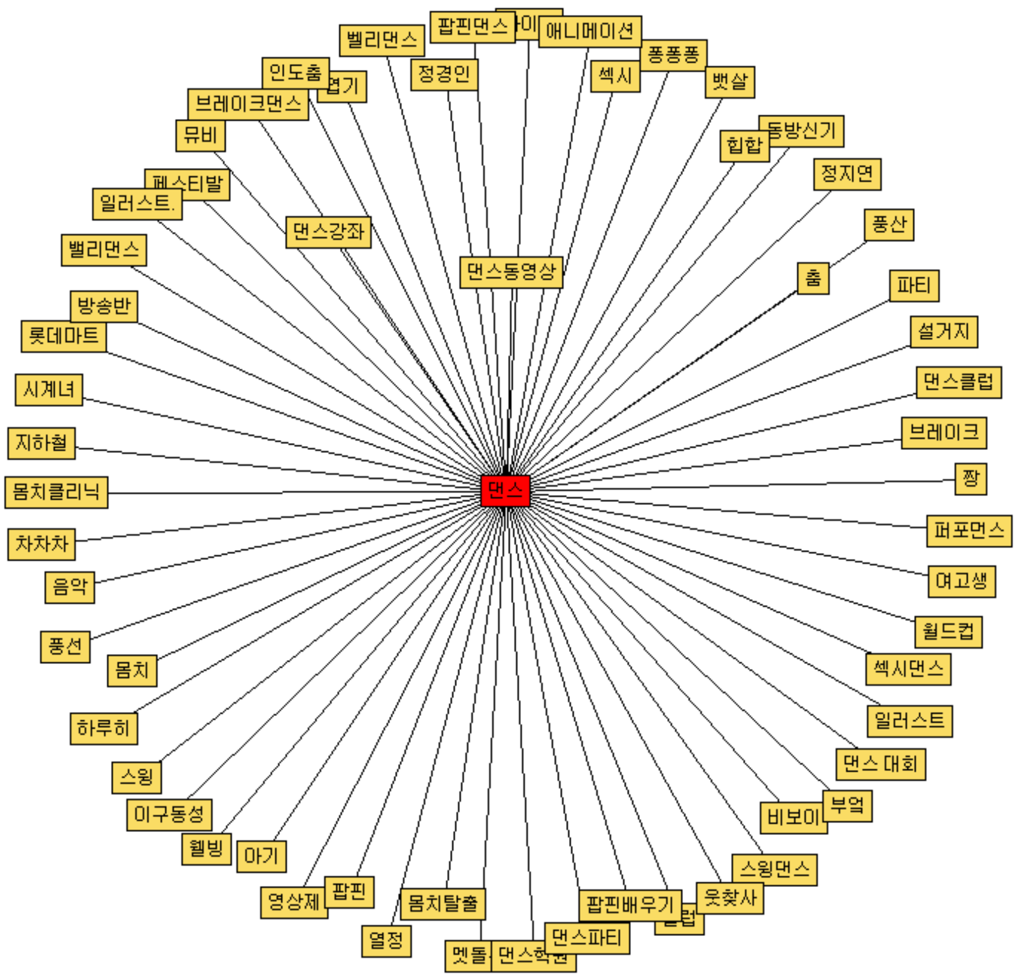
$\forall x, y \in O, \sigma(x, y) \geq 0$	positiveness
$\forall x, y, z \in O, \sigma(x, x) \geq \sigma(y, z)$	maximality
$\forall x, y \in O, \sigma(x, y) = \sigma(y, x)$	symmetry

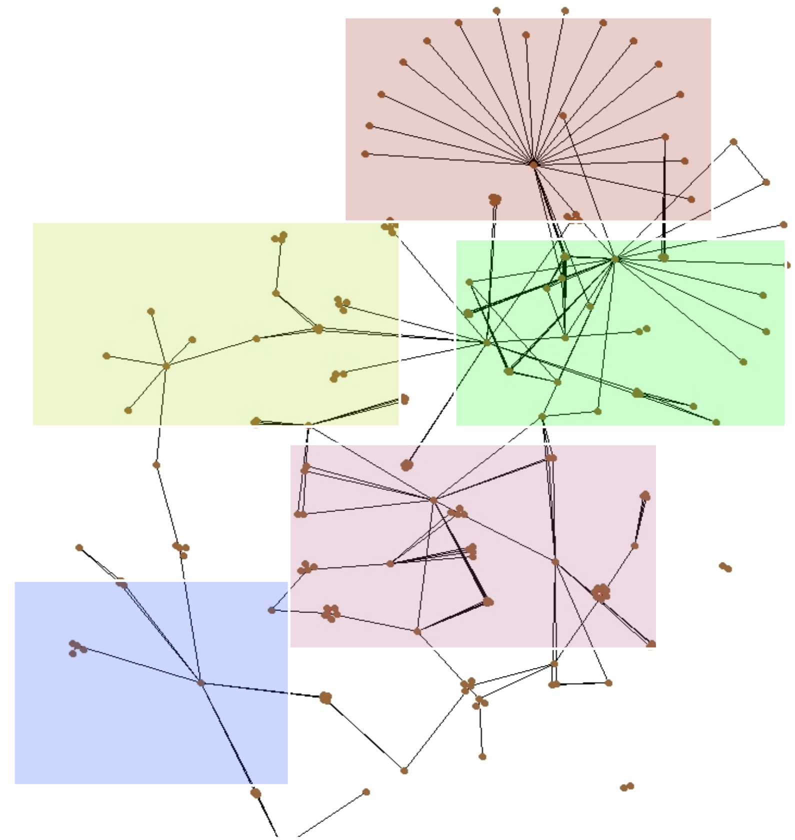
Term Similarity

$$sim_{Dice}(t_i, t_j) = \frac{2 \cdot df_{ij}}{df_i + df_j}$$

$$sim_{Goog}(\text{AION, Lineage}) = 0.1$$

Term Similarity Applications





Set Similarity

$$sim_{Jacc}(d_i, d_j) = \frac{|d_i \cap d_j|}{|d_i \cup d_j|}$$

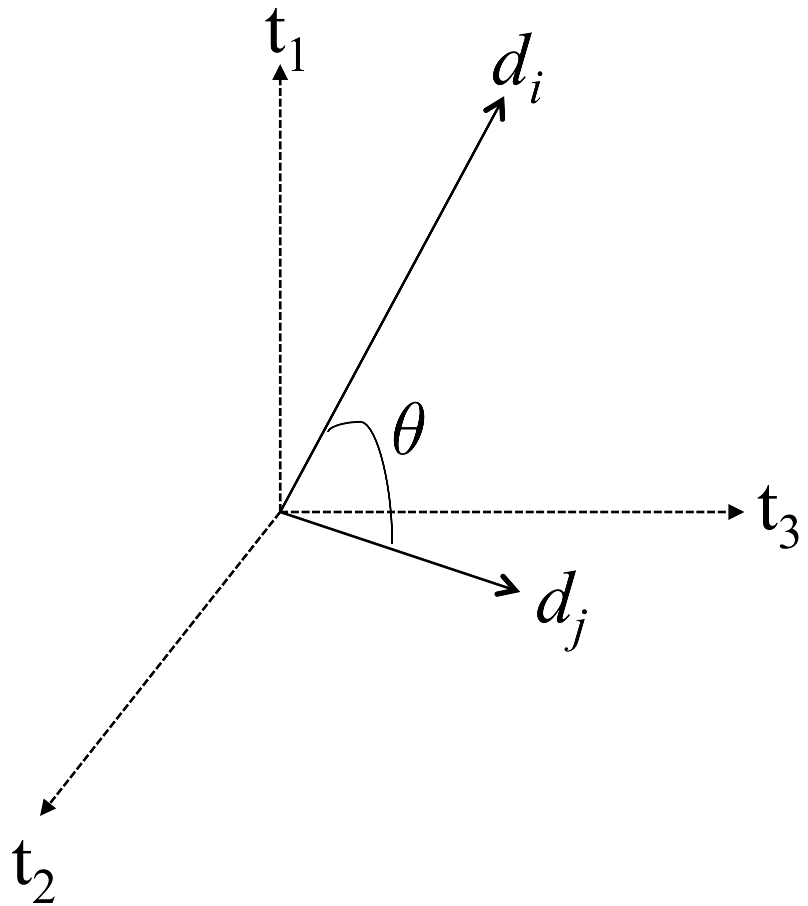
$$sim_{Dice}(d_i, d_j) = \frac{2 \times |d_i \cap d_j|}{|d_i| + |d_j|}$$

$$sim_{Overlap} = \frac{|d_i \cap d_j|}{\min(|d_i|, |d_j|)}$$

NCSoft
AION

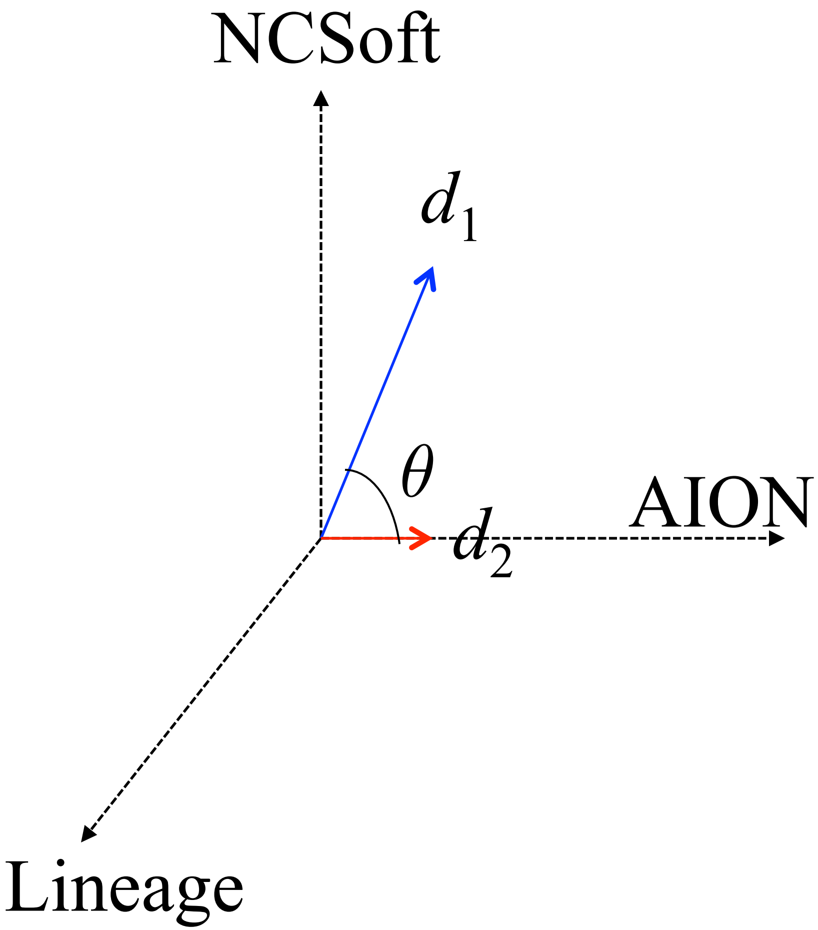
AION
Lineage

Bag Similarity – Cosine Model



$$sim_{Cos}(d_i, d_j) = \frac{\vec{d_i} \cdot \vec{d_j}}{|\vec{d_i}||\vec{d_j}|}$$

Cosine Calculation Example



d_1 = {2*NCSoft, AION}
d_2 = {AION}

$$\vec{d_1} = (2, 0, 1)$$
$$\vec{d_2} = (0, 0, 1)$$

$$sim_{Cos}(d_1, d_2)$$
$$= \frac{2 \times 0 + 0 \times 0 + 1 \times 1}{\sqrt{2^2 + 0^2 + 1^2}\sqrt{0^2 + 0^2 + 1^2}}$$
$$= \frac{1}{\sqrt{5}}$$

Bag Similarity – Pearson Correlation Coeff.

$$sim_{PCC}(d_i, d_j) = \frac{\sum_{k=1}^{n}(d_{ik} - \bar{d}_i)(d_{jk} - \bar{d}_j)}{\sqrt{\sum_{k=1}^{n}(d_{ik} - \bar{d}_i)^2 \sum_{k=1}^{n}(d_{jk} - \bar{d}_j)^2}}$$

d_1 = {2*NCSoft, AION}
d_2 = {AION}

$$\vec{d}_1 = (2, 0, 1)$$
$$\vec{d}_2 = (0, 0, 1)$$
$$\bar{d}_1 = 1$$
$$\bar{d}_2 = 1/3$$
$$sim_{PCC}(d_1, d_2) = 0$$

Collaborative Filtering

	부당거래	이층의 악당	초능력자	소셜 네트워크
서현	4	3	2	4
윤아	NA	4	5	5
유리	2	2	4	NA
수영	3	NA	5	2

$$\hat{r}_{u,v} = k \sum_{u' \in U_u} sim(u, u') \times r_{u',v}$$

$$k = 1 / \sum_{u' \in U_u} |sim(u, u')|$$

$$sim(u, u') = \frac{\sum_{v \in V_{u,u'}} (r_{u,v} - \bar{r}_u)(r_{u',v} - \bar{r}_{u'})}{\sqrt{\sum_{v \in V_{u,u'}} (r_{u,v} - \bar{r}_u)^2 \sum_{v \in V_{u,u'}} (r_{u',v} - \bar{r}_{u'})^2}}$$

Similarity of Probability Distributions

KL divergence $D(p||q) = \sum_i p_i \log \frac{p_i}{q_i}$

Information Radius (IRad) $D(p||\frac{p+q}{2}) + D(q||\frac{p+q}{2})$

L_1 norm

$$\sum_i |p_i - q_i|$$

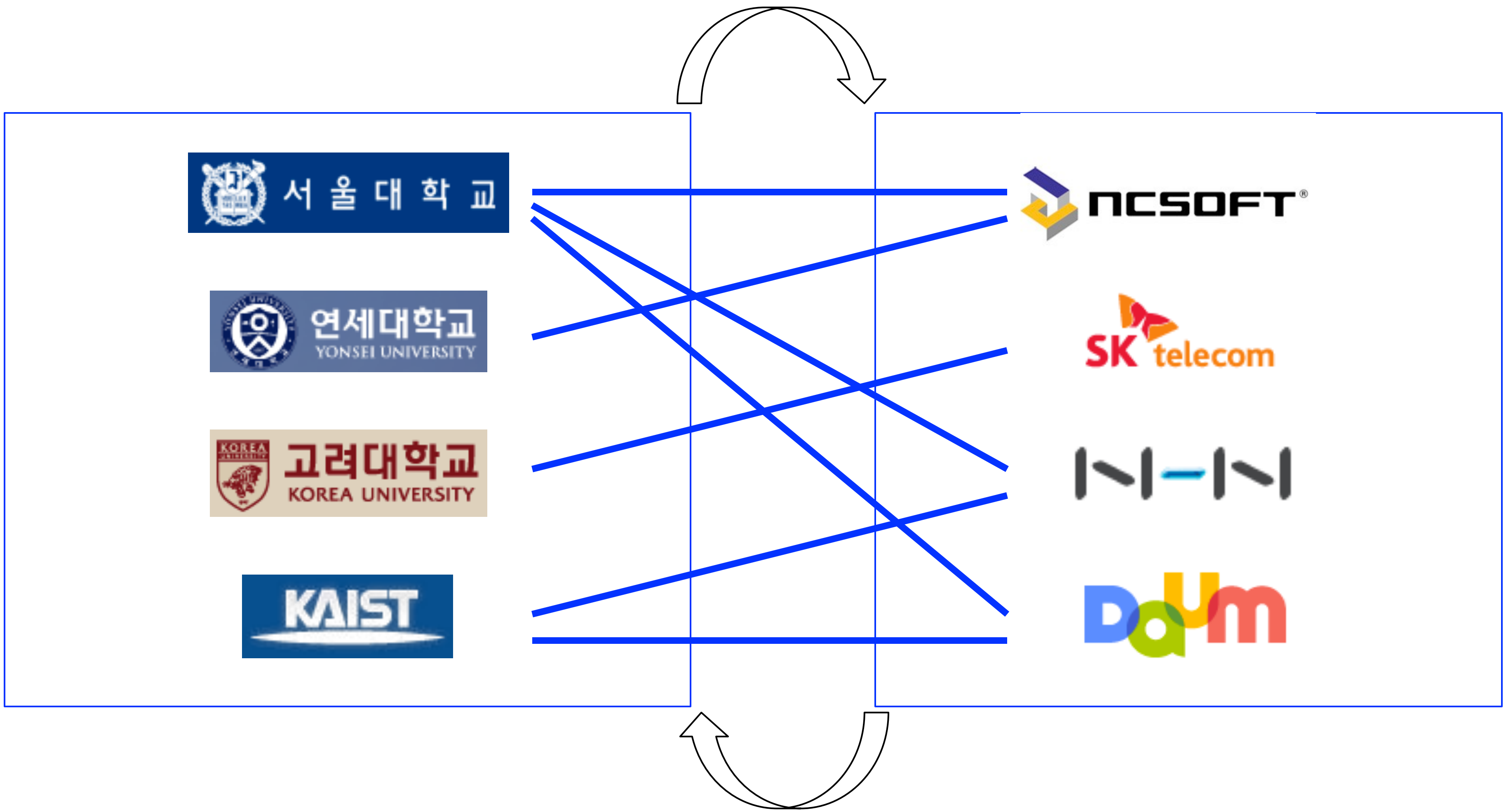
$$= \sum_i [\max(p_i, q_i) - \min(p_i, q_i)]$$

$$= \sum_i [(p_i + q_i - \min(p_i, q_i)) - \min(p_i, q_i)]$$

$$= \sum_i p_i + \sum_i q_i - 2\min(p_i, q_i)$$

$$= 2\left(1 - \sum_i \min(p_i, q_i)\right)$$

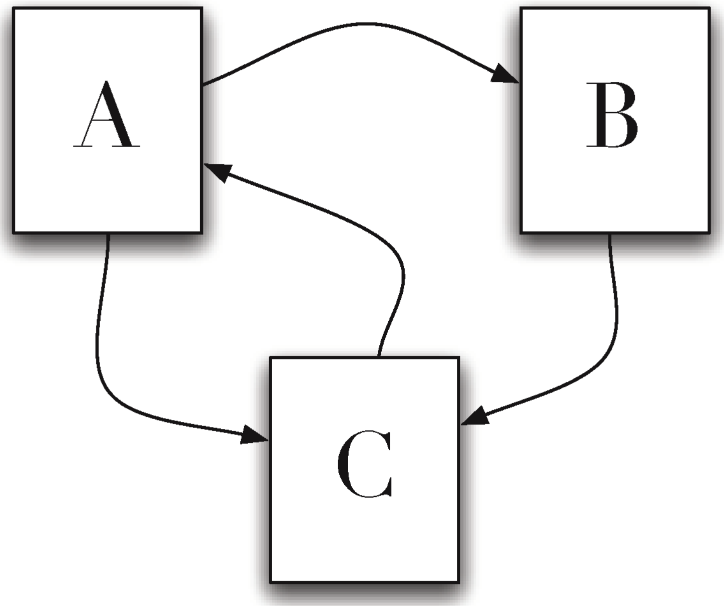
Mutual Reinforcement Principle



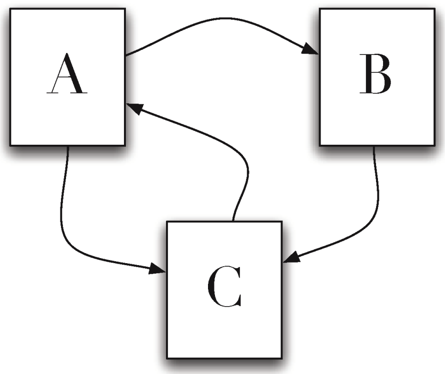
$$u(s_i) \propto \sum_{v(c_j)\sim u(s_i)} w_{ij} v(c_j) \qquad v(c_j) \propto \sum_{u(s_i)\sim v(c_j)} w_{ij} u(s_i)$$

Graph as Voting or Recommendation

Deciding the importance of a vertex within a graph



PageRank (without Random Jump)



$$PR(A) = PR(C)/1$$

$$PR(B) = PR(A)/2$$

$$PR(C) = PR(A)/2 + PR(B)/1$$

$$PR(X) = \sum_{Y \in I(X)} \frac{PR(Y)}{|O(Y)|}$$

Readings

- J. Park, B-C. Choi, and K. Kim, "A vector space approach to tag cloud similarity ranking," *Information Processing Letters*, vol. 110, 2010, pp.489–496.
- J. Park, et al., "Online Video Recommendation through Tag-Cloud Aggregation," *IEEE Multimedia*, vol. 18, no. 1, January-march 2011, pp. 78-86
- R. Swan and J. Allan, "Automatic Generation of Overview Timelines," *Proc. SIGIR*, 2000.
- G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity," *Proc. of the eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002

References

- C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.
- C. M. Grinstead and J. L. Snell, *Introduction to Probability*, 2nd Rev. Ed., American Mathematical Society, 1997.
- C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.