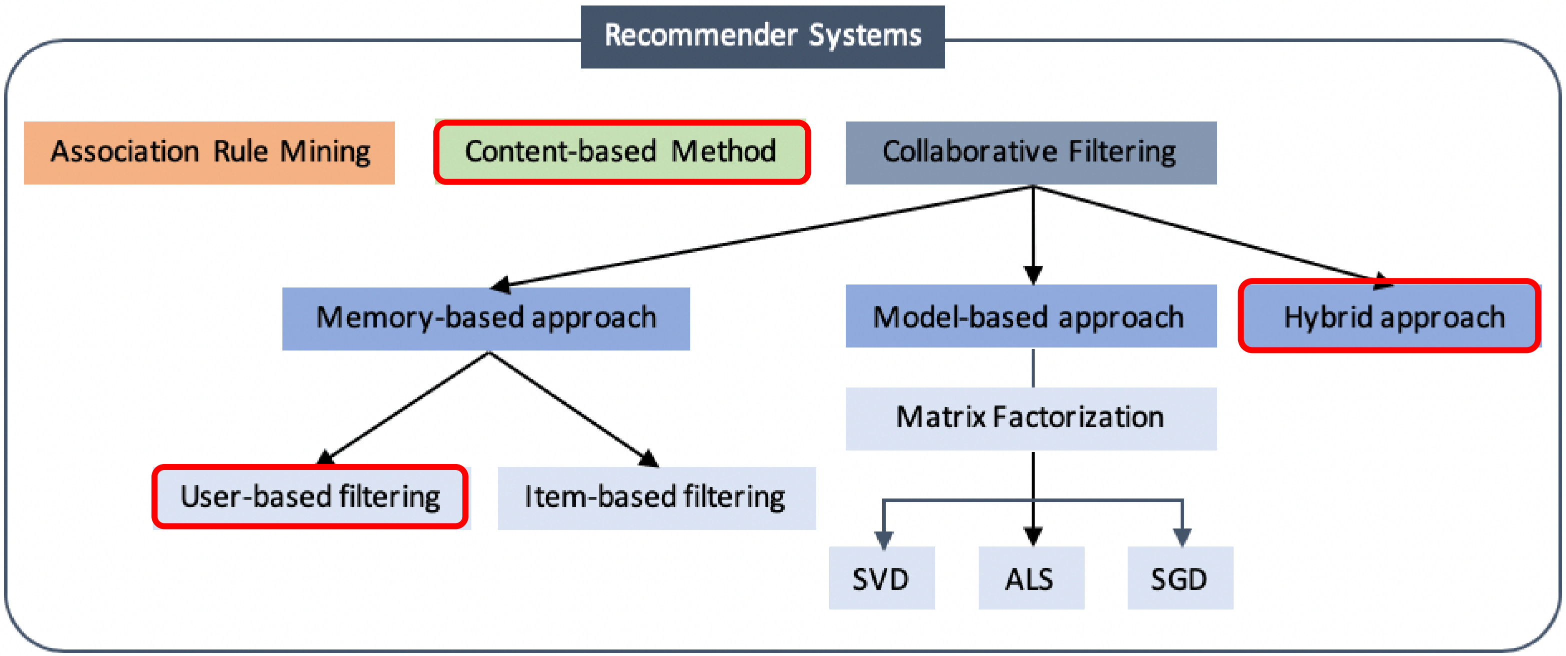
A Framework for Collaborative, Content-Based and Demographic Filtering

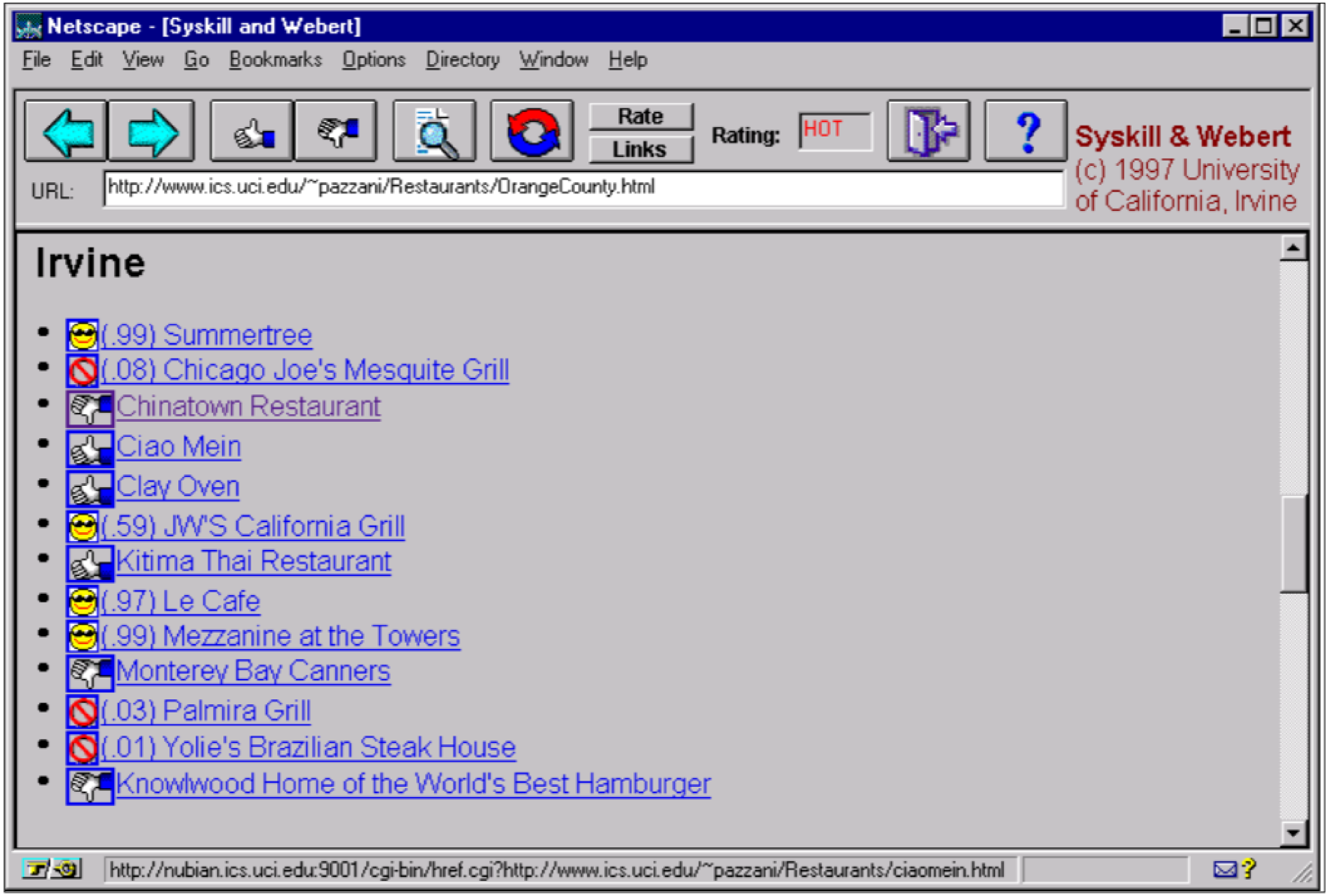
21. 04. 21

Byeong-Hyun Ko

Memory Based Collaborative Filtering



The Syskill and Webert interface



Collected Data

Table 1. Ratings of five users of five restaurants.

	Karen	Lynn	Chris	Mike	Jill
Kitima	–	+	+	+	–
Marco Polo	+	+	+	+	+
Spiga	+	–	+	–	+
Thai Touch	–	+	–	+	–
Dolce	+	–	+	–	?

- 44 Users rate restaurant with that descriptions
- Like or Not (Thumb up/down)

User Based CF

- Predict that a user might like on based **other users** who have similar taste with that of the target user

Table 1. Ratings of five users of five restaurants.

	Karen	Lynn	Chris	Mike	Jill
Kitima	–	+	+	+	–
Marco Polo	+	+	+	+	+
Spiga	+	–	+	–	+
Thai Touch	–	+	–	+	–
Dolce	+	–	+	–	?

- Correlation(similarity) method

$$r(x, y) = \frac{\sum_{d \in documents} (R_{x,d} - \bar{R}_x)(R_{y,d} - \bar{R}_y)}{\sqrt{\sum_{d \in documents} (R_{x,d} - \bar{R}_x)^2 \sum_{d \in documents} (R_{y,d} - \bar{R}_y)^2}}$$

User Based CF (con'd)

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1	1	1	1	-1
Macro Polo	1	1	1	1	1
Spiga	1	-1	1	-1	1
Thai Touch	-1	1	-1	1	-1

- Treat Pos(+) to 1, Neg(-) to -1

	Karen	Lynn	Chris	Mike	Jill
Karen	1.00000	-0.577350	0.577350	-0.577350	1.00000
Lynn	-0.57735	1.000000	-0.333333	1.000000	-0.57735
Chris	0.57735	-0.333333	1.000000	-0.333333	0.57735
Mike	-0.57735	1.000000	-0.333333	1.000000	-0.57735
Jill	1.00000	-0.577350	0.577350	-0.577350	1.00000

- Make User-User (Pearson) Correlation matrix (similarity matrix)

User Based CF (con'd)

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1	1	1	1	-1
Macro Polo	1	1	1	1	1
Spiga	1	-1	1	-1	1
Thai Touch	-1	1	-1	1	-1

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1.0	1.0	1.0	1.0	-1.0
Macro Polo	1.0	1.0	1.0	1.0	1.0
Spiga	1.0	-1.0	1.0	-1.0	1.0
Thai Touch	-1.0	1.0	-1.0	1.0	-1.0
Dolce	1.0	-1.0	1.0	-1.0	NaN

	Karen	Lynn	Chris	Mike	Jill
Karen	1.00000	-0.577350	0.577350	-0.577350	1.00000
Lynn	-0.57735	1.000000	-0.333333	1.000000	-0.57735
Chris	0.57735	-0.333333	1.000000	-0.333333	0.57735
Mike	-0.57735	1.000000	-0.333333	1.000000	-0.57735
Jill	1.00000	-0.577350	0.577350	-0.577350	1.00000

- Predict with **avg weighted sum**
  - (1+0.577+0.577+0.577)/4 = 0.682

Contents Based CF

- Recommend using item's descriptions (domain feature)
  - Feature extraction (contents analyze) is needed
  - Which words can **represent item itself**
  - *Syskill and Webert system* selects 128 most informative words
    - Using **expected informative gain**
      - *S : documents; item descriptions*

$$E(W,S) = I(S) - [p(W = present)I(S_{W=present}) + p(W = absent)I(S_{W=absent})]$$

$$I(S) = \sum_{c} -p(S_c)\log_2(p(S_c))$$

  - Learning User Profile
    - TF-IDF, Bayesian classifier -> require prespecifying the number of terms(samples) used in the profile
    - Using **Winnow algorithm**

Contents Based CF (con'd)

Table 2. The words contained in the description of 5 restaurants together with the ratings of a user for those restaurants.

	Noodle	Shrimp	Basil	Exotic	Salmon	Jill
Kitima	Y	Y	Y	Y	Y	–
Marco Polo		Y	Y			+
Spiga	Y		Y			+
Thai Touch	Y	Y		Y		–
Dolce		Y	Y		Y	?

- Informative words ; item feature ; represents from item descriptions
  - 10 words example

	Doodle	Shrimp	Basil	Exotic	Salmon	appetizer	milk	sirloin	private	farm
Kitima	1	1	1	1	0	1	1	1	0	0
Macro Polo	1	1	1	0	0	0	1	0	1	1
Spiga	1	0	0	1	1	1	0	1	1	1
Thai Touch	1	1	1	1	1	1	0	0	0	0
Dolce	0	1	0	1	0	1	1	1	0	1

Contents Based CF (con'd)

- Winnow algorithm
  - Proposed by *Littlestone and Warmuth* ;1994
  - Work like *perceptron*
  - Learning **weight(w)** with iteration
    - Threshold : decision value ; |N(x)|/2
    - X : input feature by item (each item has +/- ratings)

$$\sum w_i x_i > \tau$$

    - **Iteration rule**
      - Weights are initialized to 1
      - Finding the weighted sum
      - Above threshold && rating of X is '-' : W associate with X is divided by 2
      - Below threshold && rating of X is '+' : W associate with X is multiplied by 2

Contents Based CF (con'd)

- Winnow learning example

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1	1	1	1	-1
Macro Polo	1	1	1	1	1
Spiga	1	-1	1	-1	1
Thai Touch	-1	1	-1	1	-1

	Noodle	Shrimp	Basil	Exotic	Salmon	appetizer	milk	sirloin	private	farm
Kitima	1	1	1	1	0	1	1	1	0	0
Macro Polo	1	1	1	0	0	0	1	0	1	1
Spiga	1	0	0	1	1	1	0	1	1	1
Thai Touch	1	1	1	1	1	1	0	0	0	0
Dolce	0	1	0	1	0	1	1	1	0	1
weights	1	1	1	1	1	1	1	1	1	1

Contents Based CF (con'd)

- Winnow learning example

$$\sum w_i x_i > \tau$$

	Karen	Lynn	Chris	Mike	Jill
Kitima	$\boxed{-1}$	1	1	1	-1
Macro Polo	1	1	1	1	1
Spiga	1	-1	1	-1	1
Thai Touch	-1	1	-1	1	-1

Kitima	1	1	1	1	0	1	1	1	0	0

```
Threshold : 5
Initial Weights : [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Weighted Sum : 7
     Pos/Neg : -1
Word Vec X   : [1, 1, 1, 1, 0, 1, 1, 1, 0, 0]
Old Weights : [1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
New Weights : [0.5, 0.5, 0.5, 0.5, 1, 0.5, 0.5, 0.5, 1, 1]
```

Contents Based CF (con'd)

- Winnow learning example

$$\sum w_i x_i > \tau$$

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1.0	1.0	1.0	1.0	-1.0
Macro Polo	1.0	1.0	1.0	1.0	1.0
Spiga	1.0	-1.0	1.0	-1.0	1.0
Thai Touch	-1.0	1.0	-1.0	1.0	-1.0
Dolce	1.0	-1.0	1.0	-1.0	NaN

Spiga	1	0	0	1	1	1	0	1	1	1
Thai Touch	1	1	1	1	1	1	0	0	0	0

```
Weighted Sum : 4.0
     Pos/Neg : 1
Word Vec X   : [1, 1, 1, 0, 0, 0, 1, 0, 1, 1]
Old Weights  : [0.5, 0.5, 0.5, 0.5, 1, 0.5, 0.5, 0.5, 1, 1]
New Weights  : [1.0, 1.0, 1.0, 0.5, 1, 0.5, 1.0, 0.5, 2, 2]

Weighted Sum : 7.5
     Pos/Neg : 1
Word Vec X   : [1, 0, 0, 1, 1, 1, 0, 1, 1, 1]
Old Weights  : [1.0, 1.0, 1.0, 0.5, 1, 0.5, 1.0, 0.5, 2, 2]
New Weights  : [1.0, 1.0, 1.0, 0.5, 1, 0.5, 1.0, 0.5, 2, 2]
```

Contents Based CF (con'd)

- Winnow learning example

$$\sum w_i x_i > \tau$$

	Karen	Lynn	Chris	Mike	Jill
Kitima	-1.0	1.0	1.0	1.0	-1.0
Macro Polo	1.0	1.0	1.0	1.0	1.0
Spiga	1.0	-1.0	1.0	-1.0	1.0
Thai Touch	-1.0	1.0	-1.0	1.0	-1.0
Dolce	1.0	-1.0	1.0	-1.0	NaN

Thai Touch	1	1	1	1	1	1	0	0	0	0
Dolce	0	1	0	1	0	1	1	1	0	1

```
Weighted Sum : 5.0
     Pos/Neg : -1
Word Vec X   : [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
Old Weights  : [1.0, 1.0, 1.0, 0.5, 1, 0.5, 1.0, 0.5, 2, 2]
New Weights  : [0.5, 0.5, 0.5, 0.25, 0.5, 0.25, 1.0, 0.5, 2, 2]

Weighted Sum : 4.5
     Pos/Neg : 1
Word Vec X   : [0, 1, 0, 1, 0, 1, 1, 1, 0, 1]
Old Weights  : [0.5, 0.5, 0.5, 0.25, 0.5, 0.25, 1.0, 0.5, 2, 2]
New Weights  : [0.5, 1.0, 0.5, 0.5, 0.5, 0.5, 2.0, 1.0, 2, 4]
```

Contents Based CF (con'd)

- Winnow predict example

	Noodle	Shrimp	Basil	Exotic	Salmon	appetizer	milk	sirloin	private	farm
Kitima	1.0	1.0	1.0	1.0	0.0	1.0	1.0	1.0	0	0
Macro Polo	1.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1	1
Spiga	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	1	1
Thai Touch	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0	0
Dolce	0.0	1.0	0.0	1.0	0.0	1.0	1.0	1.0	0	1
WEIGHTS	0.5	1.0	0.5	0.5	0.5	0.5	2.0	1.0	2	4
Milano	1.0	0.0	1.0	1.0	0.0	1.0	0.0	0.0	0	1

- Weighted sum score : 6.0
- **Larger than threshold** : positive

$$\sum w_i x_i > \tau$$

Demographic Based Recommender

- Using demographic information
  - Extract data from user's home-page with text classification
  - **Trade-off** between the **quality** of the demographic information obtained and **performance**
  - On average, 57.5% of the restaurants in the top three (winnow)

	Gender	Age	Area code	Education	Employed	Dolce
Karen	F	15	714	HS	F	+
Lynn	F	17	714	HS	F	–
Chris	F	27	714	C	T	+
Mike	M	40	714	C	T	–
Jill	F	10	714	E	F	?

Collaborative via content

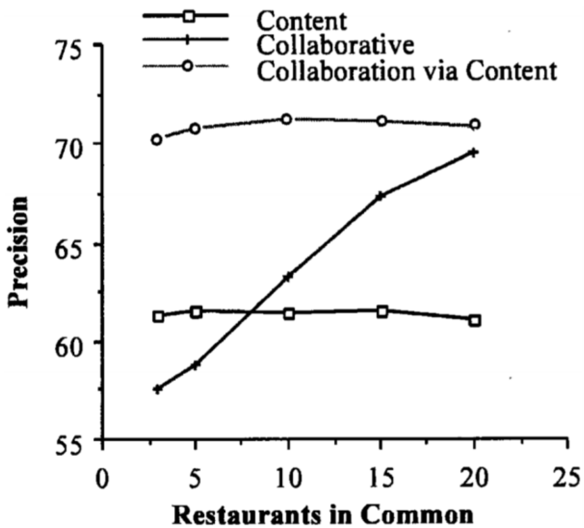
- Limitation of collaborative methods (user based)
  - Meaningful when use many rating data (cold-start problems)
  - Sparse data in real world
  - **Using contents-profile as rating matrix** (using pearson correlation)
  - On average, 70.1% of the restaurants in the top three

Table 4. Content-based profiles of five users plus their ratings for a particular restaurant

	Noodle	Shrimp	Basil	Exotic	Salmon	Dolce
Karen	2.5	0	0.2	0	0	+
Lynn	1.1	0	1.1	1.5	0	–
Chris	1.5	0	3.5	1.5	0.5	+
Mike	1.1	1.1	2.1	2.0	2.5	–
Jill	1.1	2.2	0	0	3.5	?

Combining Multiple Profiles

- Compare performance in difference environments
  - Training with data **only** in *“Southern Orange County”*
  - Test with data from **mixture** “Southern/Northern” area
  - **Content-based** : insensitive to the distribution
    - Few words referred to specific cities or geographic regions
  - **Collaborative** : poor at sparse situations
    - Performance is increased when ratings in common is raised
  - **Collaboration via content** : higher performance than others
    - Calc similarity with distribution insensitive content-based data
  - **Combining all method**
    - Sum of rank in each methods (highest to lowest : 5 to 1)
    - On average, 72.1% of the restaurants in the top three



Conclusion

Table 5. The information available for inducing a user's rating for a restaurant

Restaurants	People			Content		
	Karen	Lynn	Jill	Noodle	Shrimp	Basil
Kitima	–	+	–	Y	Y	Y
Marco Polo	+	+	+		Y	Y
Dolce	+	–	?		Y	Y
Gender	F	F	F			
Age	15	17	10			
Area code	714	714	714			
	Demographics					

- : User based CF
- : Contents based
- : Demographic based
- : Collaborative via content
- : Combining all methods

- Methods assist the user in **finding relevant information** (from data)
  - Each algorithms use **different forms** of information
  - Hybrid approaches **use more of the available information** for better results
  - Future works
    - Latent semantic indexing (using SVD for sparse problem)
    - Using continuous values