Human-Information Systems Laboratory

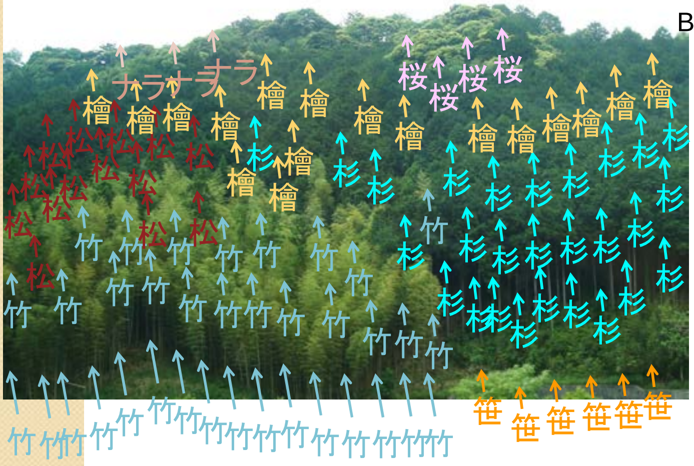
Cluster Analysis and Data Mining

Cluster Analysis is a fundamental technique for summarizing large scale data sets by partitioning similar objects into clusters. In this lecture, the basic concept of clustering is introduced with its applications.

Concept of Clustering

- How can we recognize a forest?

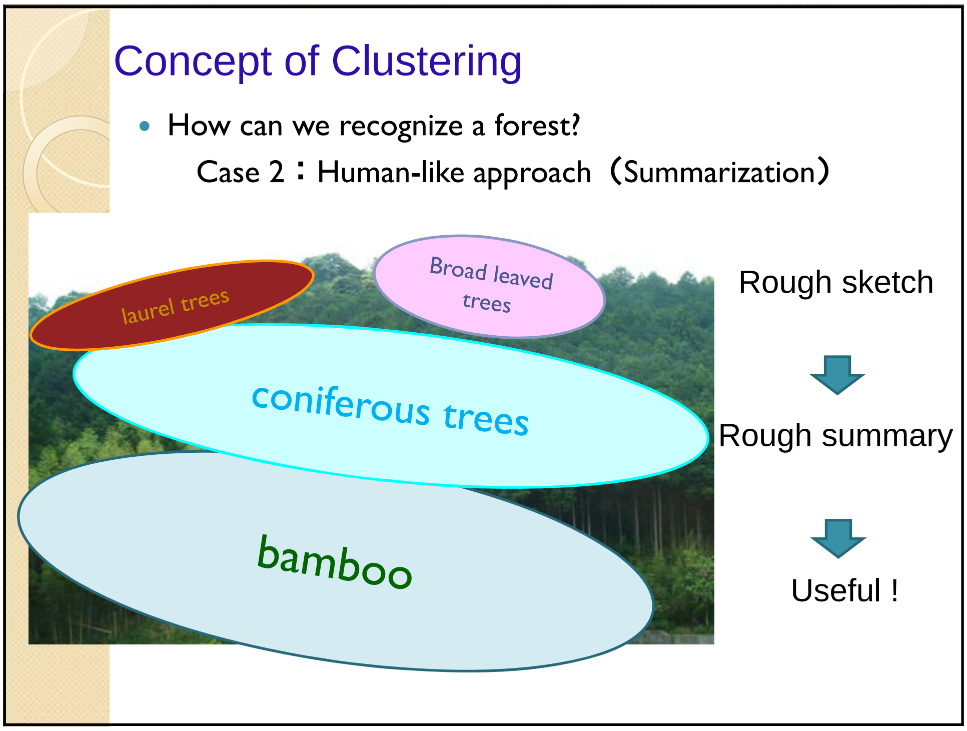
  Case I ： Machine-like approach (Enumeration)



Detailed information
But Too Complicated !

Useless !

Some people cannot see the forest for the trees.



Concept of Clustering

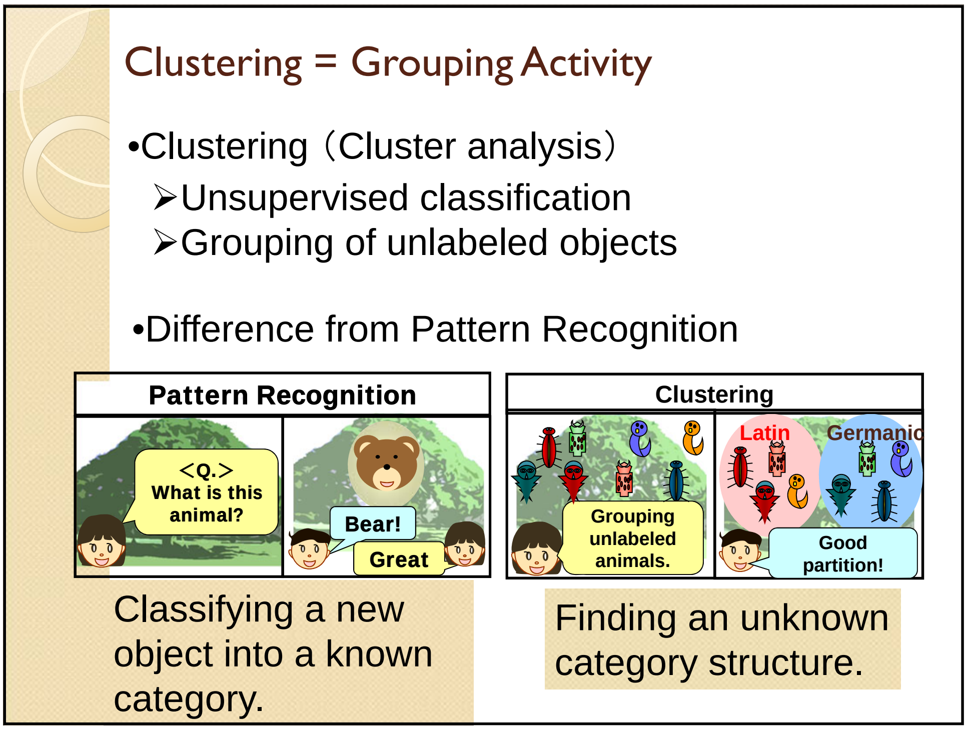
- How can we recognize a forest?

  Case 2：Human-like approach （Summarization）

Rough sketch

Rough summary

Useful !



Clustering = Grouping Activity

•Clustering（Cluster analysis）

➢Unsupervised classification

➢Grouping of unlabeled objects

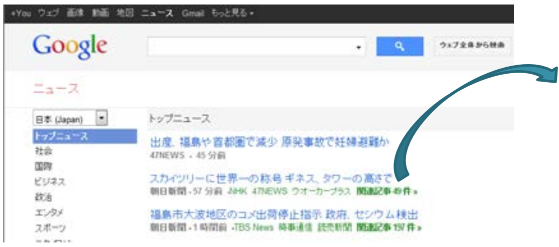
•Difference from Pattern Recognition

Classifying a new object into a known category.

Finding an unknown category structure.

Application of Clustering

- Automatic Classification in Google News



+You ウェブ 画像 動画 地図 ニュース Gmail もっと見る ▾

Google

ウェブ全体から検索

ニュース

日本 (Japan)

トップニュース
社会
国際
ビジネス
政治
エンタメ
スポーツ

トップニュース

出産、福島や首都圏で減少 原発事故で妊婦避難か
47NEWS - 45 分前

スカイツリーに世界一の称号 ギネス、タワーの高さで
朝日新聞 - 57 分前 NHK 47NEWS ウオーカープラス 関連記事 49 件 »

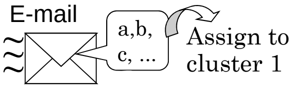
福島市大波地区のコメ出荷停止指示 政府、セシウム検出
朝日新聞 - 1 時間前 TBS News 時事通信 読売新聞 関連記事 197 件 »

Automatic search of similar news documents

- Automatic E-mail Classification by document clustering

Keywords

Document	a	b	c	d	e	f	g	
1	✓	✓	✓					Cluster 1
2	✓		✓		✓			
3		✓	✓					
4	✓	✓					✓	
5				✓		✓	✓	Cluster 2
6		✓			✓	✓		
7				✓	✓		✓	



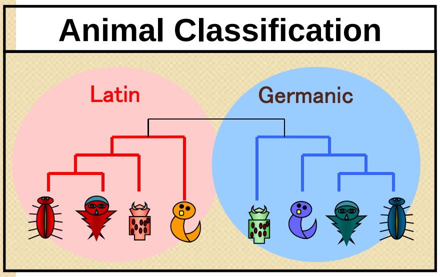
Clustering is a hot topic!

Variants of Clustering

•Hierarchical Method

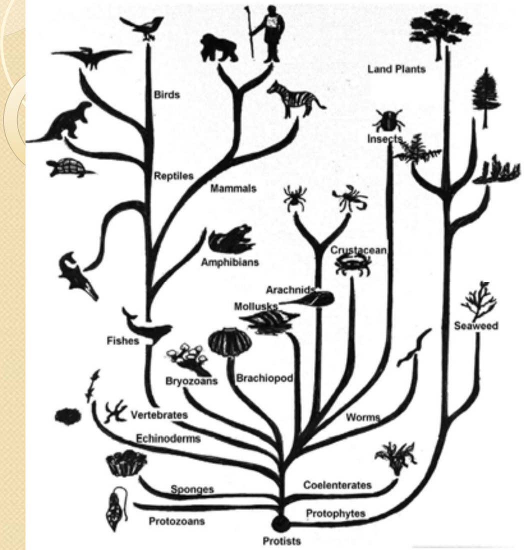
➢Gradually increasing (decreasing) the number of clusters.

➢Summary by Tree-like structure



Animal Classification

- Various clusters with various cutting levels
- Computationally expensive for large data sets



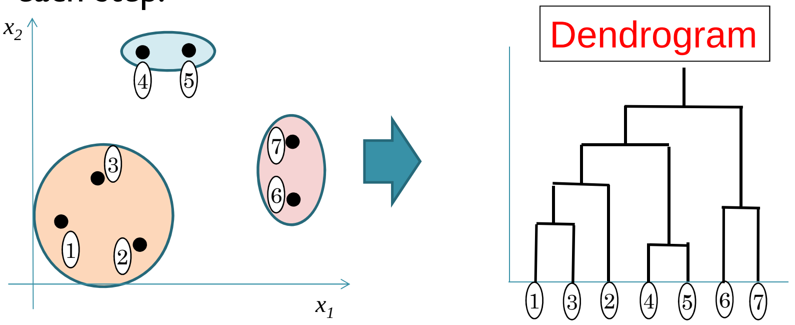
A sample of “Tree of Life Gallery”

Hierarchical clustering cannot applied to larger data sets.

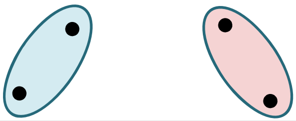
Batch process by non-hierarchical method is preferred.

Aggregation Process

- All disjoint objects are merged into clusters one by one such that most familiar clusters are merged in each step.



- What are the `most familiar clusters’?

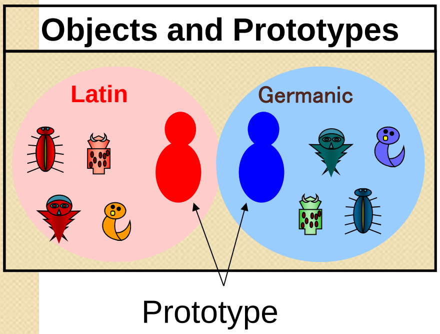


Various distances between multi-object clusters.

Variants of Clustering

•Non-hierarchical Method

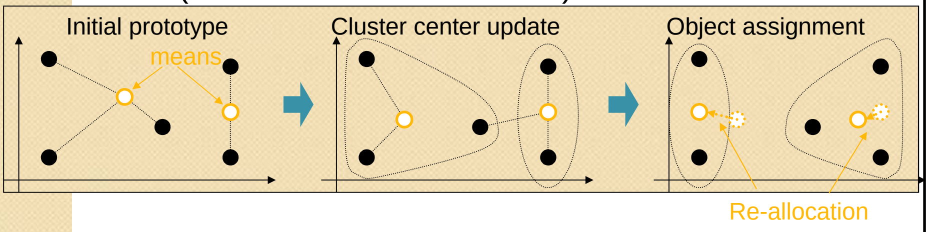
➢Extracting a pre-defined (fixed) number of clusters, each of which represented by a cluster prototype.

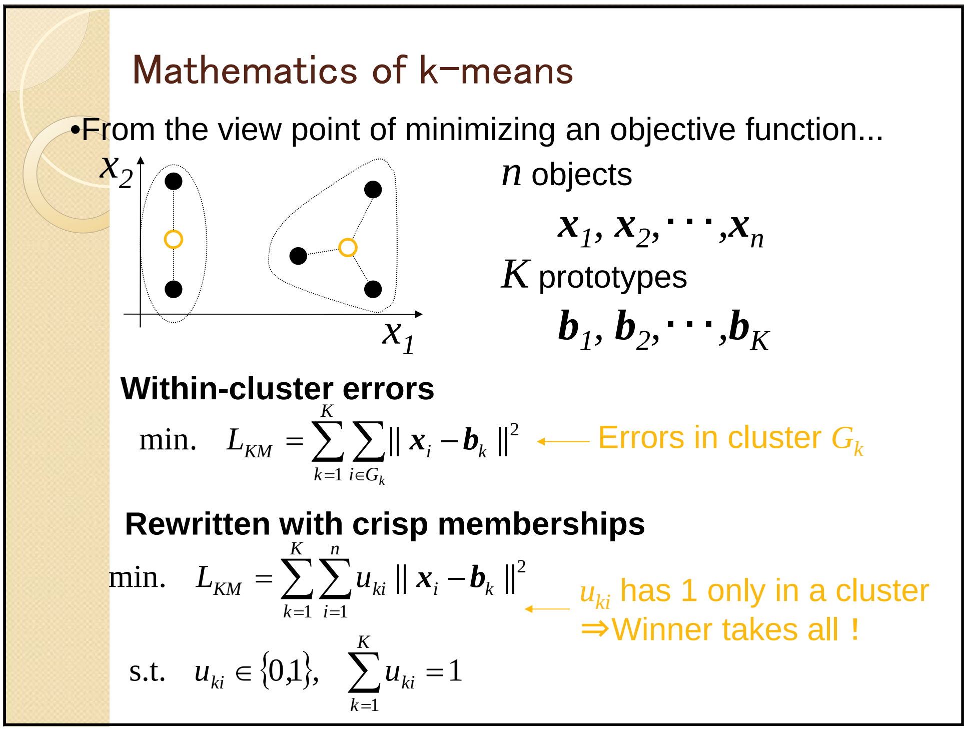


- Minimizing the distances between objects and prototypes
- Suitable for large data sets
- Various extensions with different prototypes

Basic method

- k-means clustering algorithm [MacQueen 1967]
  - Iterative algorithm composed of two phases
  - Prototype estimation phase (Cluster center update)
  - Nearest prototype assignment （Cluster membership update）
  - Prototype = k mean vector (k is the number of clusters)





Mathematics of k-means

- From the view point of minimizing an objective function...

n objects

$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$

K prototypes

$\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_K$

Within-cluster errors

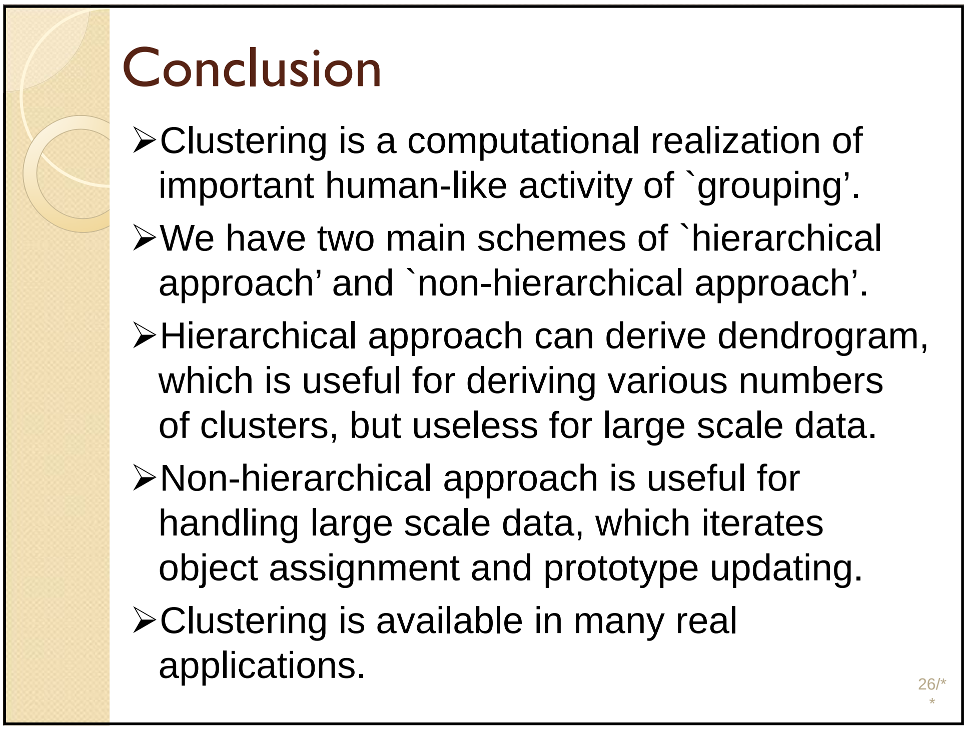
$$\min. \quad L_{KM} = \sum_{k=1}^{K} \sum_{i \in G_k} \| \boldsymbol{x}_i - \boldsymbol{b}_k \|^2 \longleftarrow \text{Errors in cluster } G_k$$

Rewritten with crisp memberships

$$\min. \quad L_{KM} = \sum_{k=1}^{K} \sum_{i=1}^{n} u_{ki} \| \boldsymbol{x}_i - \boldsymbol{b}_k \|^2 \longleftarrow$$

u_{ki} has 1 only in a cluster
⇒Winner takes all !

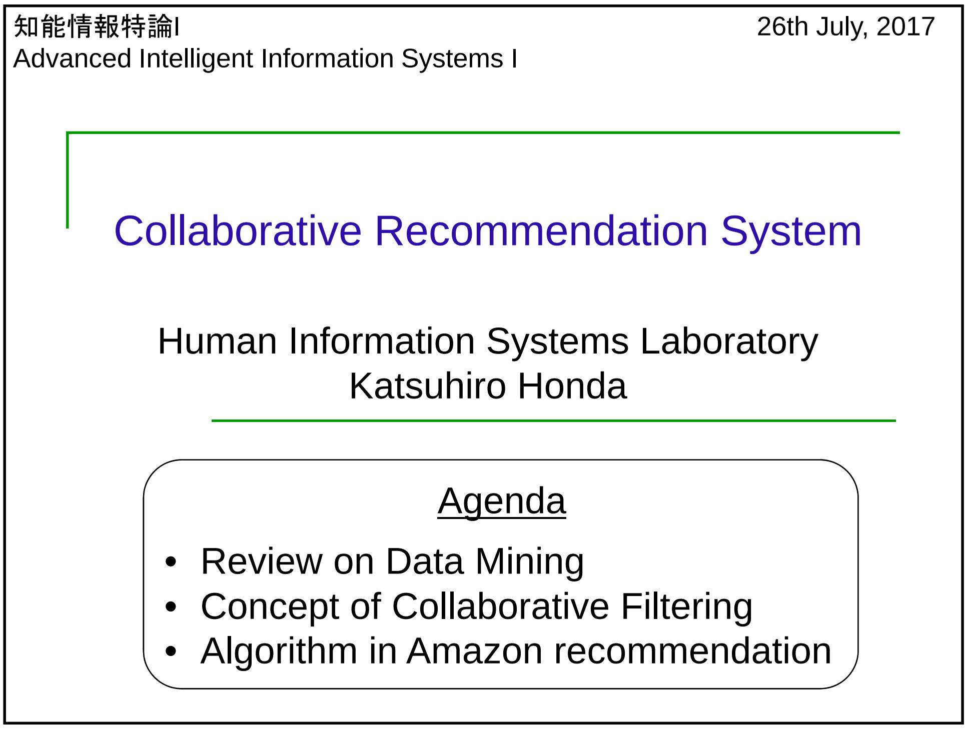
$$\text{s.t.} \quad u_{ki} \in \{0,1\}, \quad \sum_{k=1}^{K} u_{ki} = 1$$



Conclusion

- ➢Clustering is a computational realization of important human-like activity of `grouping'.
- ➢We have two main schemes of `hierarchical approach' and `non-hierarchical approach'.
- ➢Hierarchical approach can derive dendrogram, which is useful for deriving various numbers of clusters, but useless for large scale data.
- ➢Non-hierarchical approach is useful for handling large scale data, which iterates object assignment and prototype updating.
- ➢Clustering is available in many real applications.

26/*
*



知能情報特論I
Advanced Intelligent Information Systems I

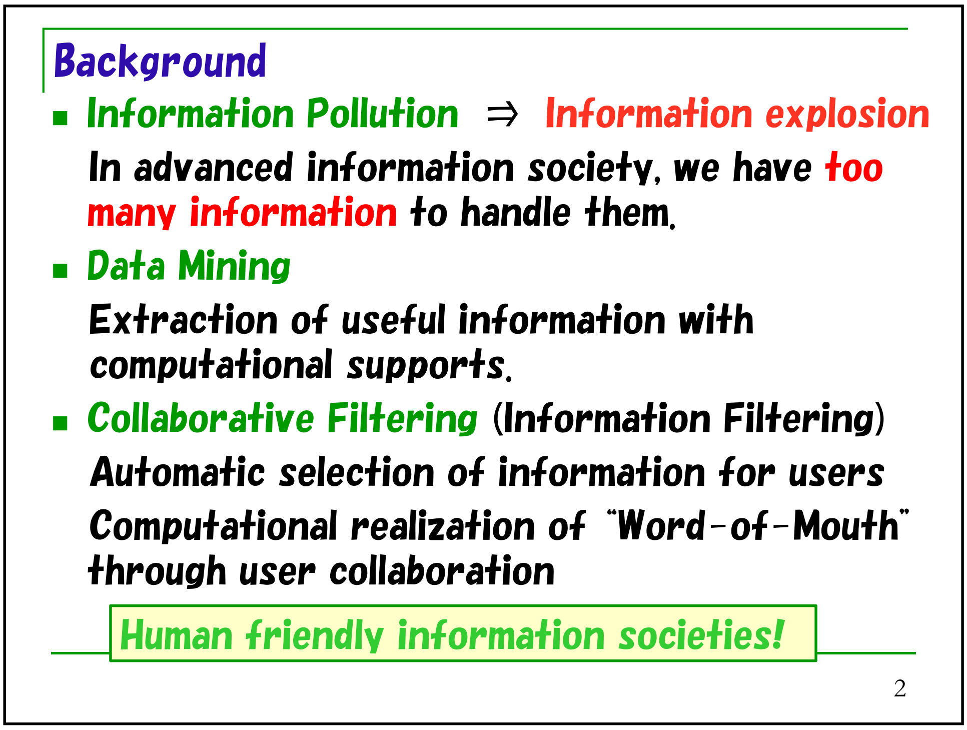
26th July, 2017

Collaborative Recommendation System

Human Information Systems Laboratory
Katsuhiro Honda

Agenda

- Review on Data Mining
- Concept of Collaborative Filtering
- Algorithm in Amazon recommendation



Background

- **Information Pollution** ⇒ **Information explosion**

  In advanced information society, we have too many information to handle them.

- **Data Mining**

  Extraction of useful information with computational supports.

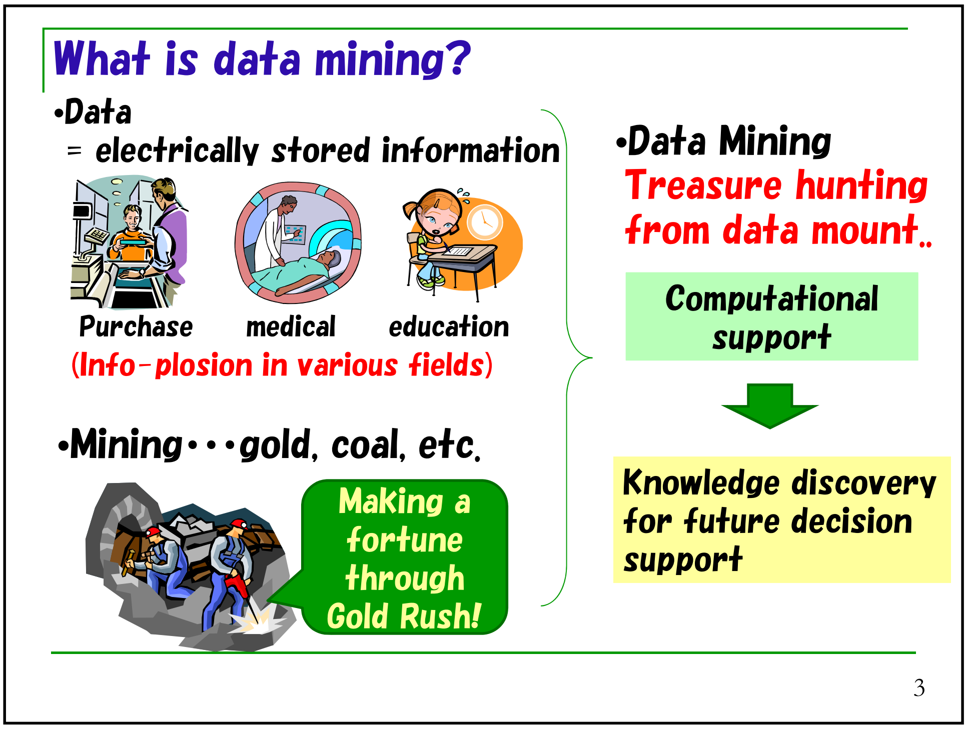
- **Collaborative Filtering** **(Information Filtering)**

  Automatic selection of information for users

  Computational realization of "Word-of-Mouth" through user collaboration

Human friendly information societies!

2



What is data mining?

- Data
  = electrically stored information

Purchase medical education

(Info-plosion in various fields)

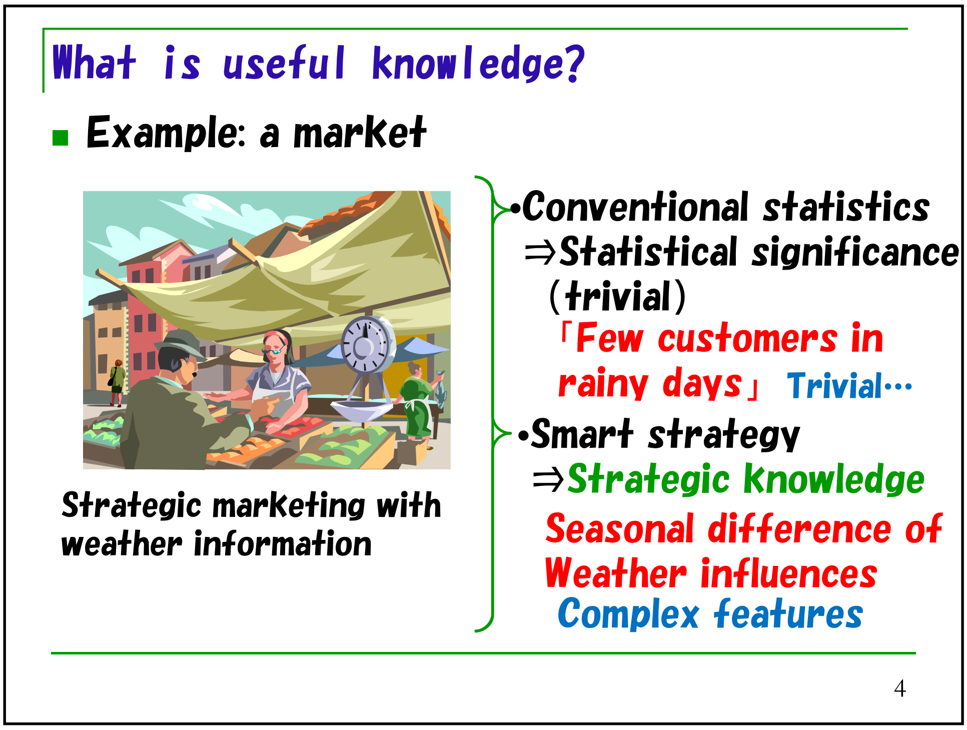
- Mining･･･gold, coal, etc.

Making a fortune through Gold Rush!

- Data Mining
  Treasure hunting from data mount..

Computational support

Knowledge discovery for future decision support



What is useful knowledge?

- Example: a market

Strategic marketing with weather information

- Conventional statistics
  ⇒Statistical significance
  (trivial)
  「Few customers in rainy days」 Trivial…
- Smart strategy
  ⇒Strategic knowledge
  Seasonal difference of Weather influences
  Complex features

3 key elements of Data Mining

- Observation・・・We must carefully observe data.
- Sampling・・・We should extract meaningful group.

  Clustering! (My early lecture)

  Extension from conventional statistics
- Correlation ・・・We can find pattern in the group.

  (correlation (association) rule analysis)

  Utilization of conventional statistics !

For Human-friendly information society…

- IT method for supporting finding preferable information from info-mountain

Here is Wally!

Not everyone searching for Wally!

An example...

How do you find YOUR interesting information from data masses?

I have nothing to do today.
Let' s enjoy a "Funny" video content.
How can we find a "Funny" one ?

Limit of conventional information retrieval

- Ranking: Popularity-based Recommendation

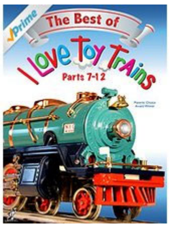
  Majority tendency = Ignoring personal feelings

  Unable to recommend based on personal preferences

- Keyword search: Content-based recommendation
  - Comparison between keyword and content label

  Label may not fit user' s feeling.

  Funny ≠ Feeling for Content



How will you find a funny content in real world?

- **Word-of-Mouth Communication**
  - You often prefer to ask your friends for promising recommendation!
  - You often recommend favorite items to your friends!

Why?

You expect that your friend have similar preference to you!

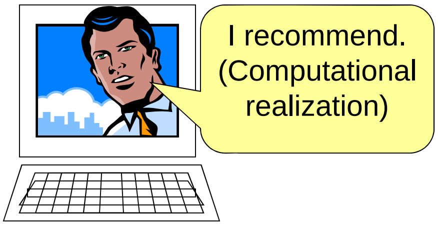
Word-of-Mouth communication
= Recommendation from Friends!

Collaborative Filtering
= Preference-based information filtering

- **Considering how users feel for items**

  ⇒Comparison of users feelings such as ratings on items

Personalized recommendation is achieved by finding `Friends'!



Recommendation by neighborhood user search

Collaborative filtering: Virtual Word-of-Mouth

Advantages of collaborative filtering

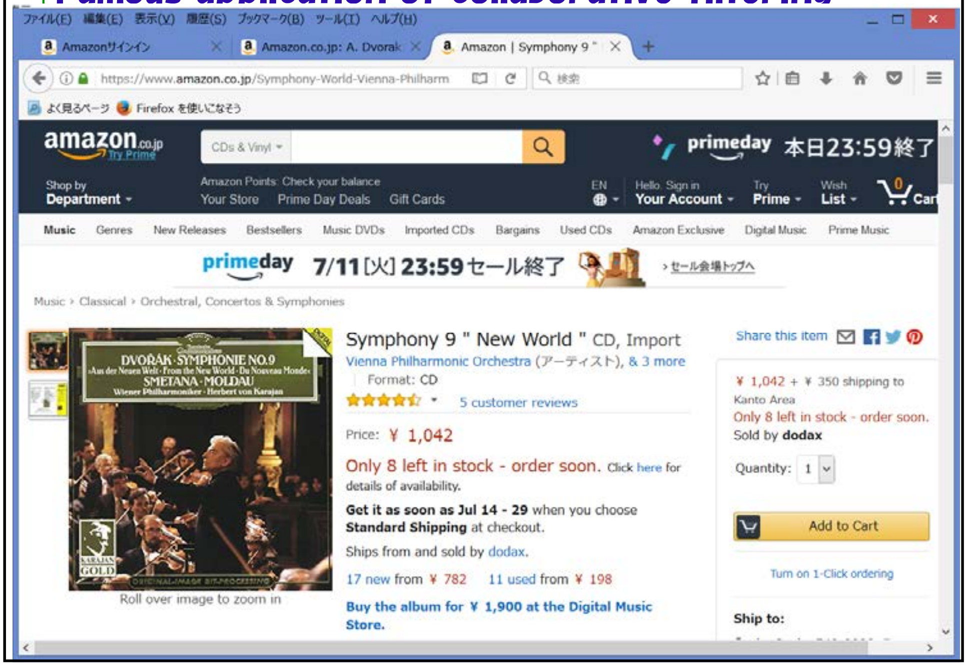
(i) **Available without content analysis**
We cannot check all labels of all contents.

(ii) **Recommend based on item quality or tastes**
We need not present appropriate keywords.

(iii) **Provide serendipitous recommendation**
We can find (unknown) valuable content.

16

Famous application of collaborative filtering



Symphony 9 " New World " CD, Import
Vienna Philharmonic Orchestra (アーティスト), & 3 more
Format: CD
5 customer reviews

Price: ¥ 1,042

Only 8 left in stock - order soon. Click here for details of availability.

Get it as soon as Jul 14 - 29 when you choose **Standard Shipping** at checkout.

Ships from and sold by dodax.

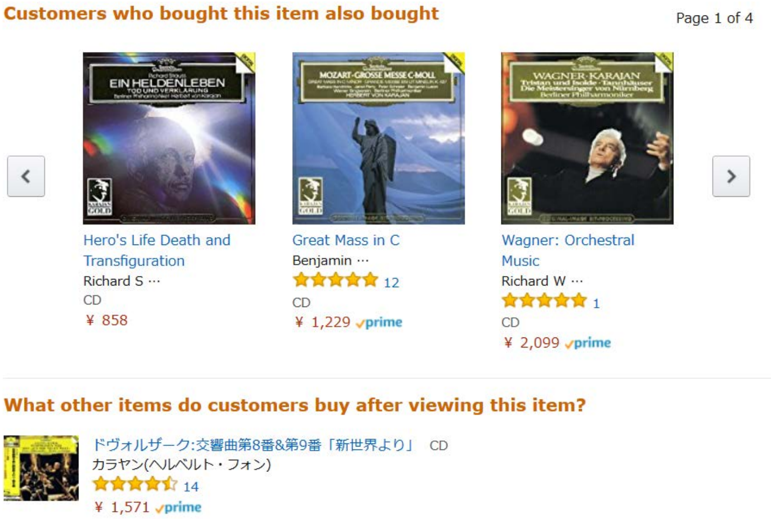
17 new from ¥ 782   11 used from ¥ 198

Buy the album for ¥ 1,900 at the Digital Music Store.

Roll over image to zoom in

¥ 1,042 + ¥ 350 shipping to Kanto Area
Only 8 left in stock - order soon.
Sold by **dodax**

Information filtering and E-commerce



Customers who bought this item also bought Page 1 of 4

Hero's Life Death and Transfiguration
Richard S …
CD
¥ 858

Great Mass in C
Benjamin …
★★★★★ 12
CD
¥ 1,229 ✓prime

Wagner: Orchestral Music
Richard W …
★★★★★ 1
CD
¥ 2,099 ✓prime

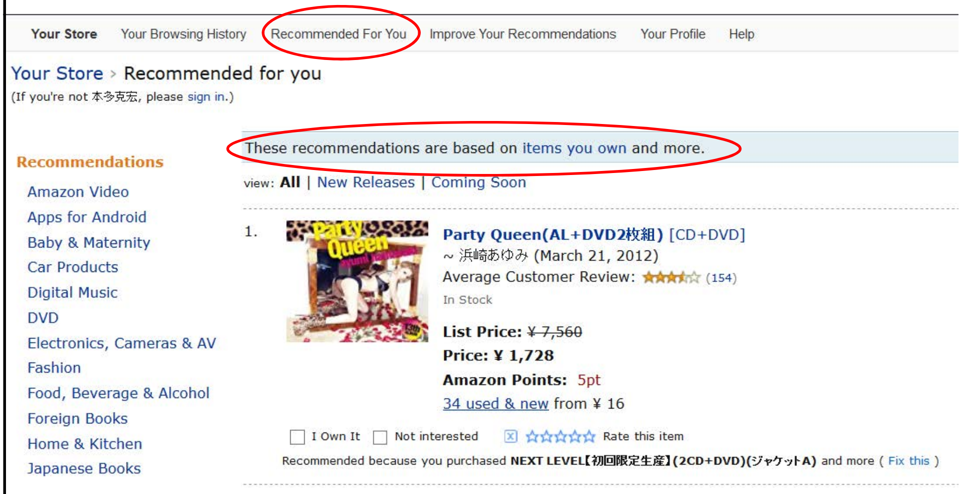
What other items do customers buy after viewing this item?

ドヴォルザーク:交響曲第8番&第9番「新世界より」 CD
カラヤン(ヘルベルト・フォン)
★★★★☆ 14
¥ 1,571 ✓prime

Efficient marketing based on user preferences

18

Personalized Recommendation in Amazon



Your Store › Recommended for you
(If you're not 本多克宏, please sign in.)

Recommendations
- Amazon Video
- Apps for Android
- Baby & Maternity
- Car Products
- Digital Music
- DVD
- Electronics, Cameras & AV
- Fashion
- Food, Beverage & Alcohol
- Foreign Books
- Home & Kitchen
- Japanese Books

These recommendations are based on items you own and more.

view: **All** | New Releases | Coming Soon

1. **Party Queen(AL+DVD2枚組)** [CD+DVD]
~ 浜崎あゆみ (March 21, 2012)
Average Customer Review: ★★★★☆ (154)
In Stock

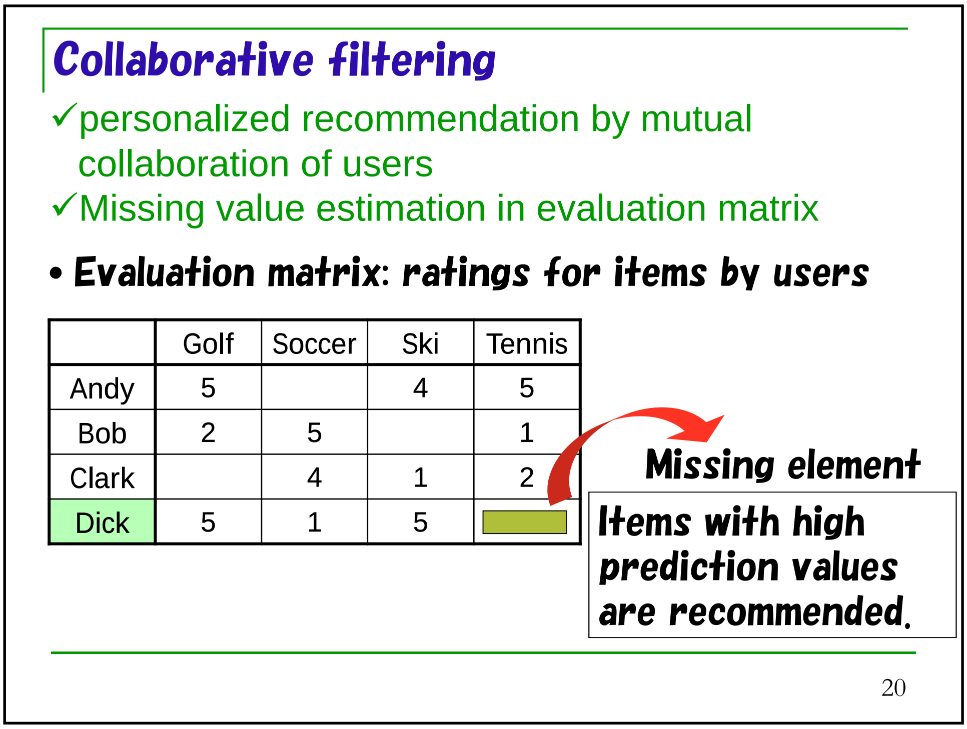
List Price: ~~¥ 7,560~~
Price: ¥ 1,728
Amazon Points: 5pt
34 used & new from ¥ 16

□ I Own It □ Not interested ☒ ☆☆☆☆☆ Rate this item

Recommended because you purchased **NEXT LEVEL【初回限定生産】(2CD+DVD)(ジャケットA)** and more (Fix this)

Personalized recommendation based on my personal purchase history

19



Collaborative filtering

✓personalized recommendation by mutual collaboration of users

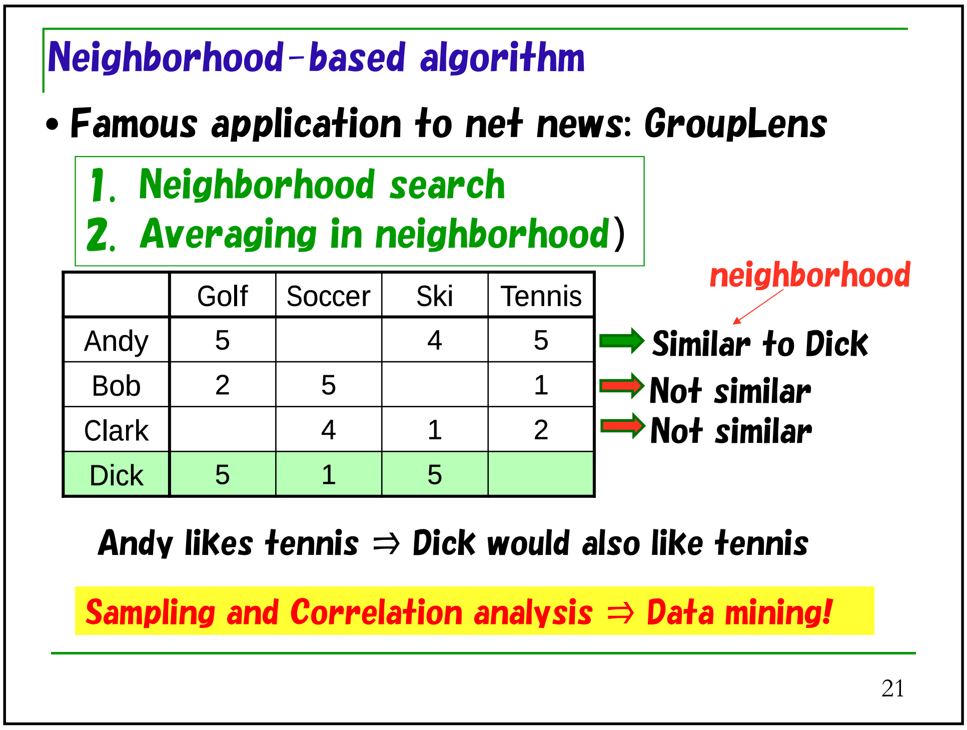
✓Missing value estimation in evaluation matrix

- **Evaluation matrix: ratings for items by users**

	Golf	Soccer	Ski	Tennis
Andy	5		4	5
Bob	2	5		1
Clark		4	1	2
Dick	5	1	5	🟩

Missing element

Items with high prediction values are recommended.



Neighborhood-based algorithm

- **Famous application to net news: GroupLens**

1. Neighborhood search
2. Averaging in neighborhood)

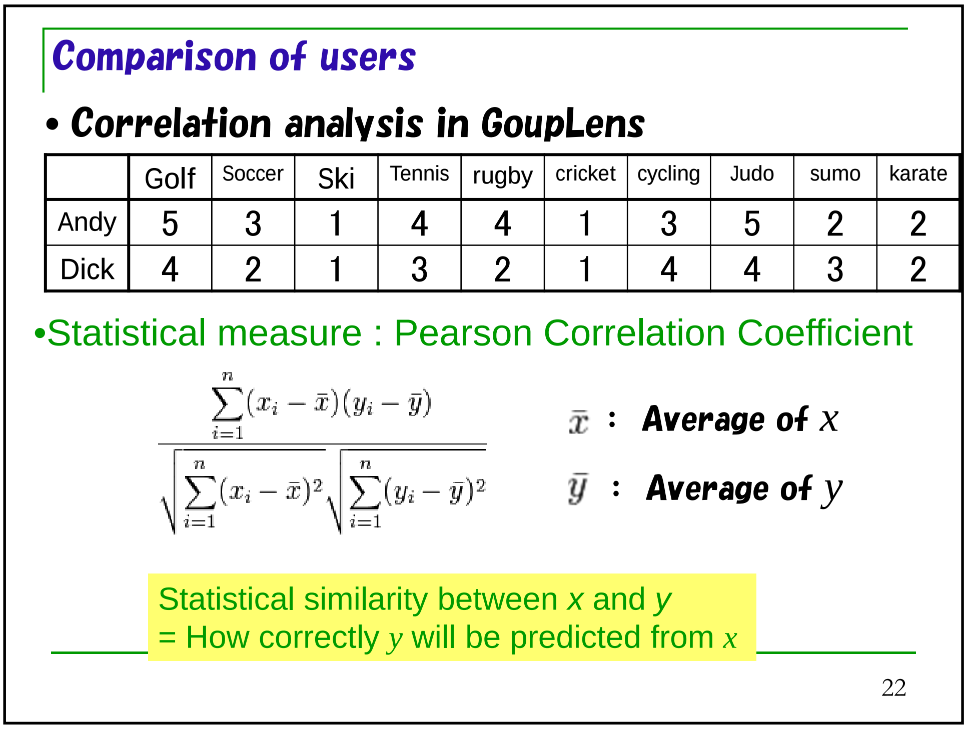
	Golf	Soccer	Ski	Tennis
Andy	5		4	5
Bob	2	5		1
Clark		4	1	2
Dick	5	1	5	

neighborhood

➡ Similar to Dick
➡ Not similar
➡ Not similar

Andy likes tennis ⇒ Dick would also like tennis

Sampling and Correlation analysis ⇒ Data mining!



Comparison of users

- **Correlation analysis in GoupLens**

	Golf	Soccer	Ski	Tennis	rugby	cricket	cycling	Judo	sumo	karate
Andy	5	3	1	4	4	1	3	5	2	2
Dick	4	2	1	3	2	1	4	4	3	2

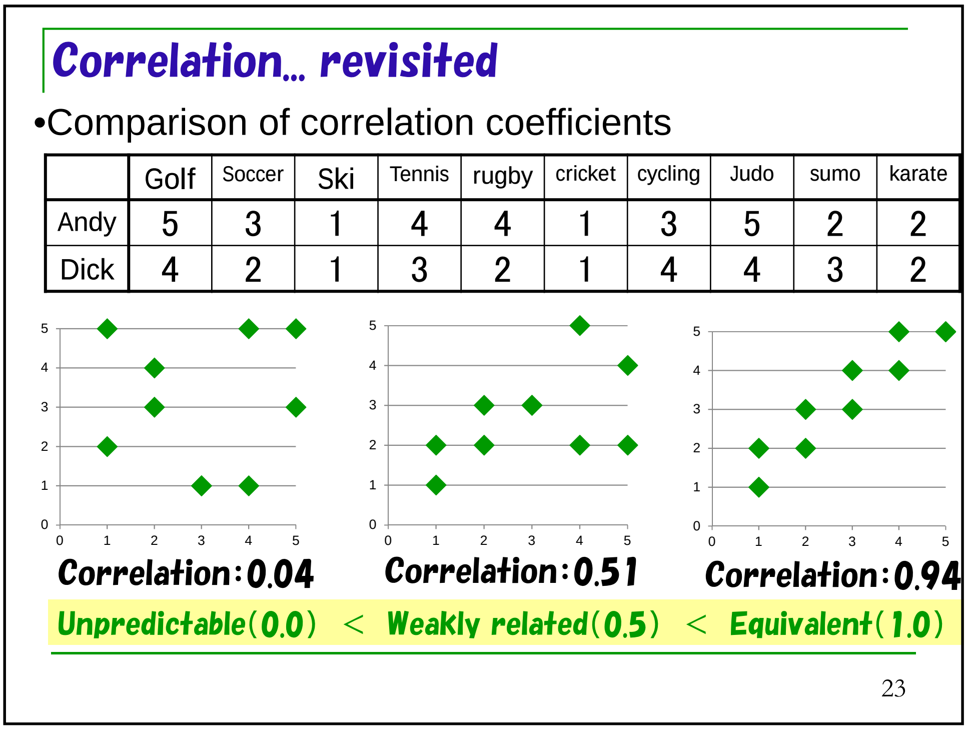
•Statistical measure : Pearson Correlation Coefficient

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$\bar{x}$: **Average of** x

$\bar{y}$: **Average of** y

Statistical similarity between *x* and *y*
= How correctly *y* will be predicted from *x*



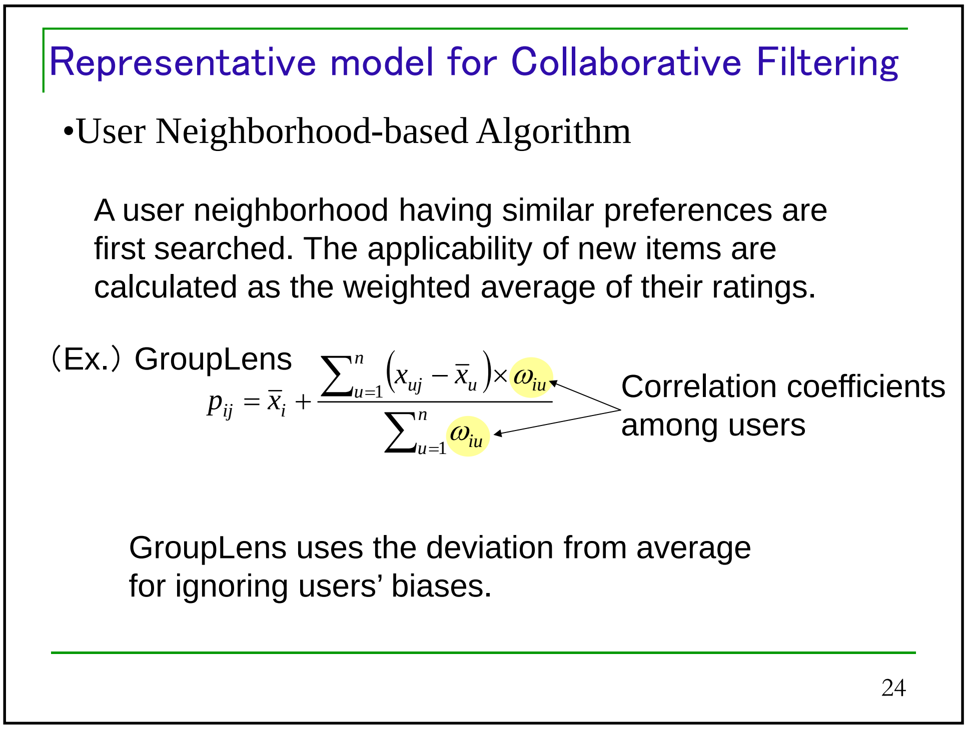
Correlation… revisited

•Comparison of correlation coefficients

	Golf	Soccer	Ski	Tennis	rugby	cricket	cycling	Judo	sumo	karate
Andy	5	3	1	4	4	1	3	5	2	2
Dick	4	2	1	3	2	1	4	4	3	2

Correlation:0.04 **Correlation:0.51** **Correlation:0.94**

Unpredictable(0.0) < Weakly related(0.5) < Equivalent(1.0)



Representative model for Collaborative Filtering

•User Neighborhood-based Algorithm

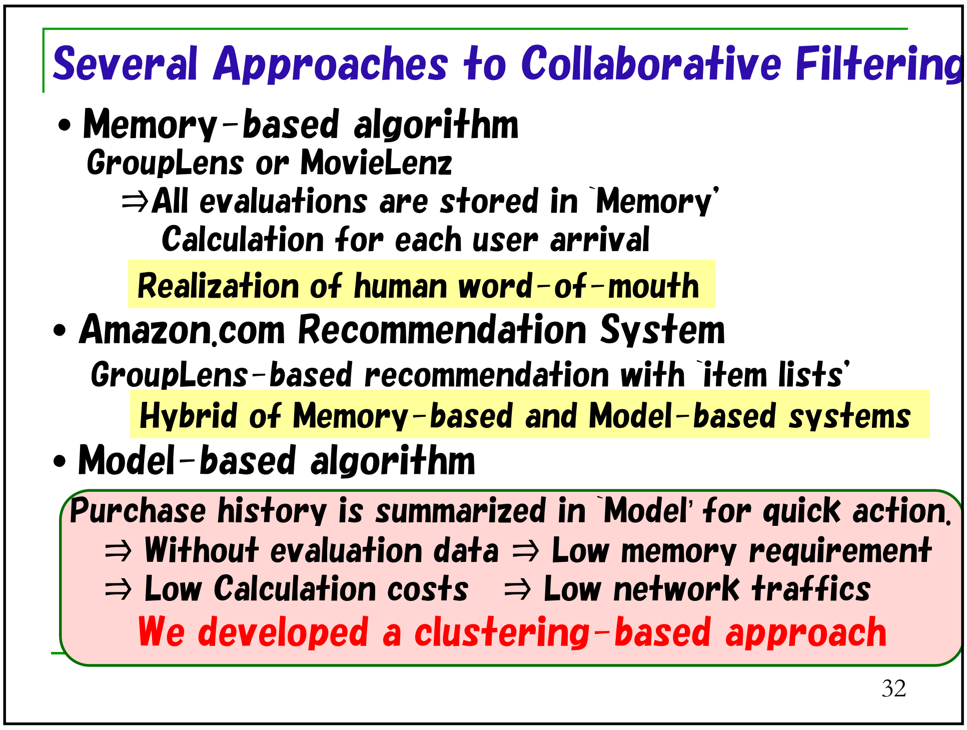
A user neighborhood having similar preferences are first searched. The applicability of new items are calculated as the weighted average of their ratings.

（Ex.）GroupLens

$$p_{ij} = \bar{x}_i + \frac{\sum_{u=1}^{n}\left(x_{uj} - \bar{x}_u\right)\times\omega_{iu}}{\sum_{u=1}^{n}\omega_{iu}}$$

Correlation coefficients among users

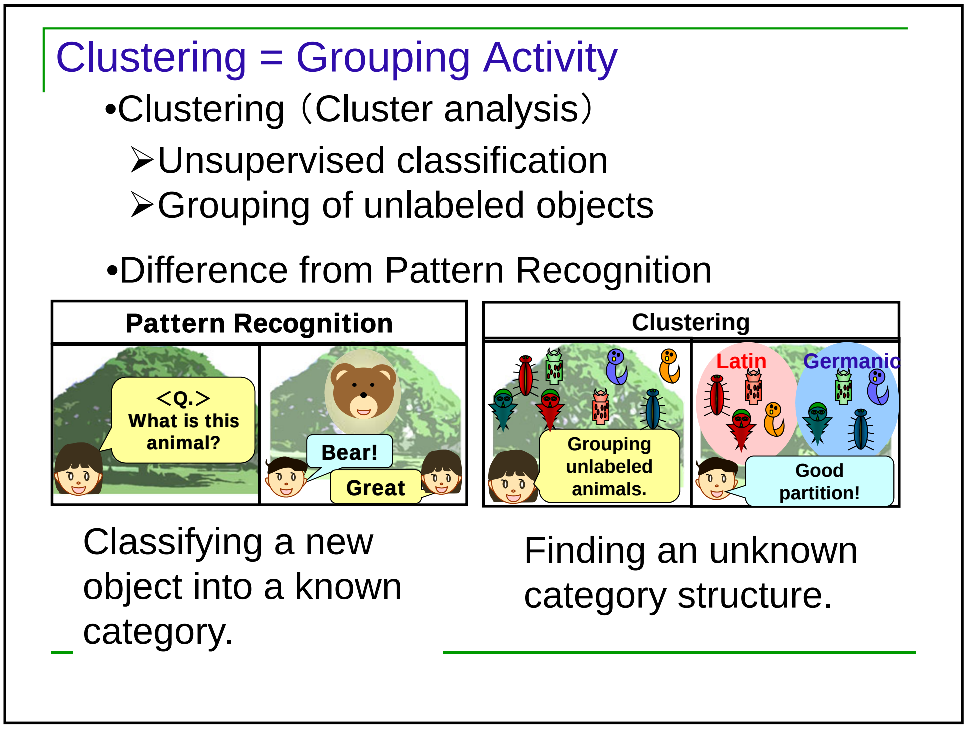
GroupLens uses the deviation from average for ignoring users' biases.



Several Approaches to Collaborative Filtering

- **Memory-based algorithm**
  GroupLens or MovieLenz
  ⇒All evaluations are stored in `Memory'
  Calculation for each user arrival
  Realization of human word-of-mouth
- **Amazon.com Recommendation System**
  GroupLens-based recommendation with `item lists'
  Hybrid of Memory-based and Model-based systems
- **Model-based algorithm**

Purchase history is summarized in `Model' for quick action.
⇒ Without evaluation data ⇒ Low memory requirement
⇒ Low Calculation costs ⇒ Low network traffics
We developed a clustering-based approach



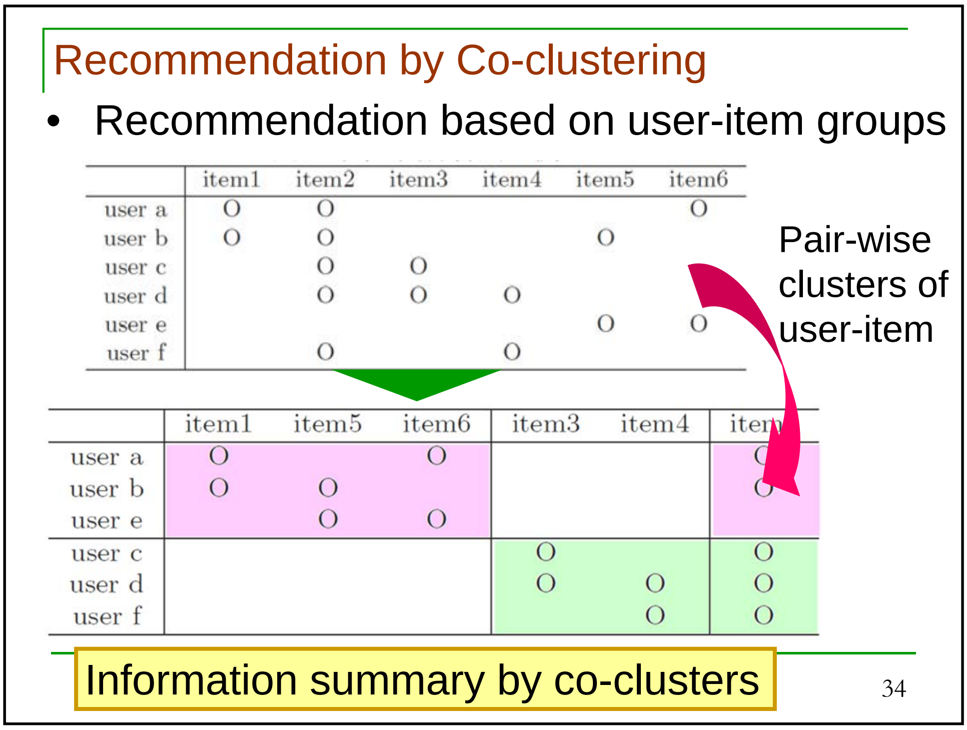
Clustering = Grouping Activity

- •Clustering（Cluster analysis）
  - ➢Unsupervised classification
  - ➢Grouping of unlabeled objects
- •Difference from Pattern Recognition

Pattern Recognition
<Q.> What is this animal?
Bear!
Great
Clustering
Grouping unlabeled animals.
Latin
Germanic
Good partition!

Classifying a new object into a known category.

Finding an unknown category structure.



Recommendation by Co-clustering

- Recommendation based on user-item groups

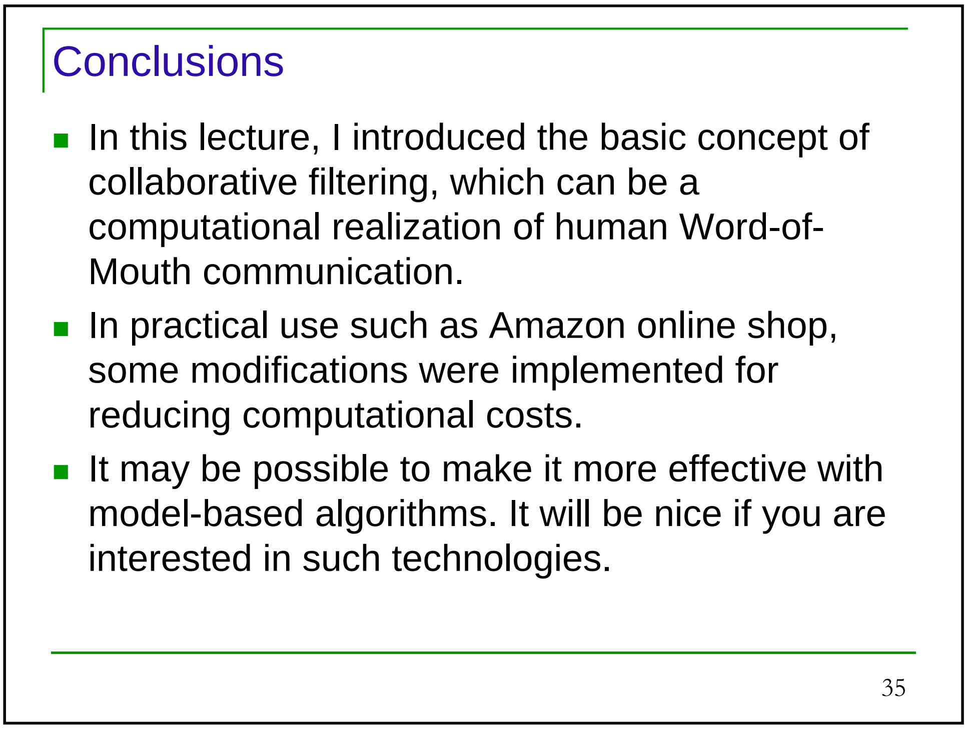
	item1	item2	item3	item4	item5	item6
user a	O	O				O
user b	O	O			O	
user c		O	O			
user d		O	O	O		
user e					O	O
user f		O		O		

	item1	item5	item6	item3	item4	item
user a	O		O			O
user b	O	O				O
user e		O	O			
user c				O		O
user d				O	O	O
user f					O	O

Pair-wise clusters of user-item

Information summary by co-clusters

34



Conclusions

- In this lecture, I introduced the basic concept of collaborative filtering, which can be a computational realization of human Word-of-Mouth communication.
- In practical use such as Amazon online shop, some modifications were implemented for reducing computational costs.
- It may be possible to make it more effective with model-based algorithms. It will be nice if you are interested in such technologies.