大規模データ処理システム
Big Data Processing System

協調フィルタリング (1)
Collaborative Filtering (1)

- 「Aを買った人はBを買っています。」
- 消費者の消費行動をログとして蓄積している。
  - 慶應太郎, 学問のすすめ, 2014/06/09, ...
  - 慶應太郎, 福翁自伝, 2014/04/30, ...
  - 慶應花子, 福翁自伝, 2014/03/31, ...
  - 湘南次郎, 尊王論, 2013/03/01, ...
  - 慶應花子, 尊王論, 2012/12/01, …

- A people, who buy A, buy B too.
- A system collects purchasing log.
  - Taro KEIO, Gakumon no susume, 2014/06/09, ...
  - Taro KEIO, Fukuou jiden, 2014/04/30, ...
  - Hanako KEIO, Fukuou jiden, 2014/03/31, ...
  - Jiro SHONAN, Sonnouron, 2013/03/01, ...
  - Hanako KEIO, Sonnouron, 2012/12/01, ...

協調フィルタリング (2)
Collaborative Filtering (2)

- アイテム毎に集計する（ユーザのベクトルを作る）。
  - 学問のすすめ: 慶應太郎
  - 福翁自伝: 慶應太郎, 慶應花子
  - 尊王論: 湘南次郎, 慶應花子

- Totaling by each item
  - Gakumon no susume: Taro KEIO
  - Fukuou jiden: Taro KEIO, Hanako KEIO
  - Sonnou ron: Jiro SHONAN, Hanako KEIO

協調フィルタリング (3)
Collaborative Filtering (3)

- アイテムごとのベクトルの類似度を計算する。
  - コサイン類似度:
    cos(a,b)　= a・b / (|a| ・ |b|)
    = 両方アクセスしたユーザ数 / a, bそれぞれにアクセスしたユーザ数
    - 1に近ければ類似している。0に近ければ類似していない。
    - cos(学問のすすめ, 福翁自伝) = 1 / (√1 * √2) = 0.707

- Make similarity of vectors of each item
  - Cosine similarity:
    cos(a,b)　= a・b / (|a|・|b|)
    = num of users who access both / num of users who access a or b
    - If similarity is close to 1, it might be similar. If similarity is close to 0, it might be not similar.
    - cos(“Gakumon no susume”, “Fukuou jiden”) = 1 / (√1 * √2) = 0.707

協調フィルタリングを考える (4)
Collaborative Filtering (4)

大きく2段階の処理

- アイテム毎に集計する
  - mapによって、<アイテム, ユーザ> というkey-valueペアを作る。
  - reduceによってアイテム毎にユーザのベクトル<アイテム, (ユーザ1, ユーザ2…)>を作成する。
- 類似度を計算する
  - cos(学問のすすめ, 慶應太郎)を計算する。
  - アイテム数^{2}の類似度が生成される。

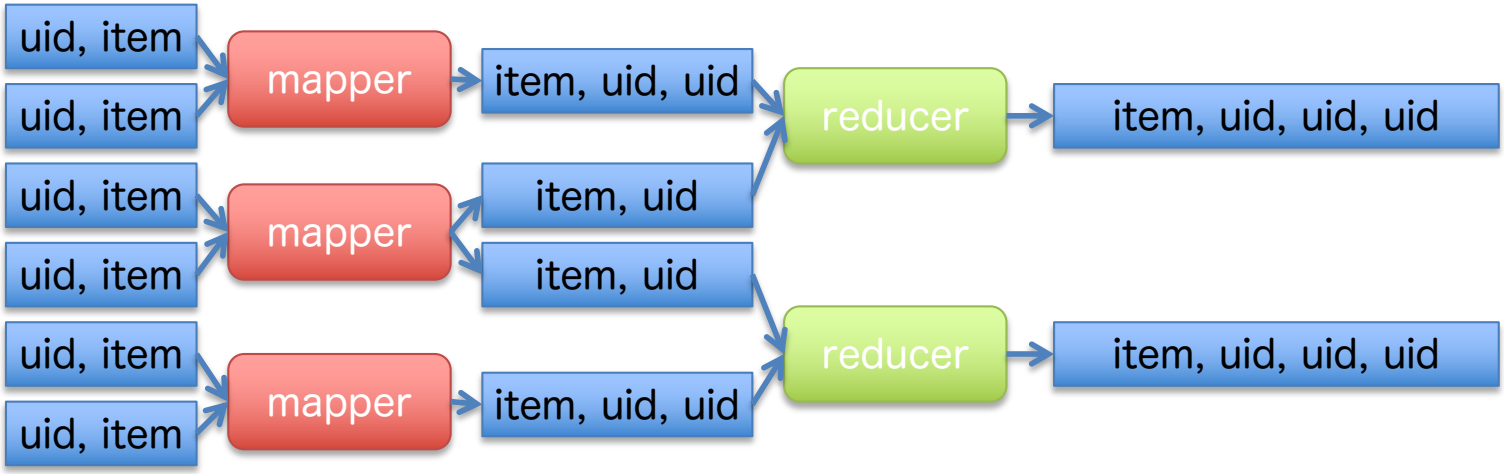
Two phases

- Totaling by items
  - Make <item, user> key-value pair using map procedure
  - Make <item, (user1, user2, ...)> vector using reduce procedure
- Make similarity
  - Make cos(item1, item2)
  - The process produces (num of items)^2 similarity

協調フィルタリングを考える (5)
Collaborative Filtering (5)

Two phases

- Totaling by items
  - Make <item, user> key-value pair using map procedure
  - Make <item, (user1, user2, ...)> vector using reduce procedure



- Make similarity
  - Make cos(item1, item2)
  - The process produces (num of items)^2 similarity

全てのアイテム間の関係を計算しなければならず、ローカリティが保てない
All combination of items must be calculated.
It means there's no locality.

協調フィルタリングを考える (6)
Collaborative Filtering (6)

- コサイン類似度を求めるためには(a, b)のアイテムペアに対する、下記の情報がわかればよい。
  - 両方を購入した人数
  - aを購入した人数
  - bを購入した人数

- アイテム毎の購入者数(aを購入した人数、bを購入した人数)はアイテムリストを作った段階で自明。
- 故に、両方を購入した人数を求めることができれば、コサイン類似度は求められる。

※コサイン類似度:
cos(a,b)= a・b / (|a| ・ |b|)
= 両方アクセスしたユーザ数 / a, bそれぞれにアクセスしたユーザ数

協調フィルタリングを考える (6)
Collaborative Filtering (6)

- When you make Cosine similarity, you need following information for a pair of a and b.
  - the number of peoples who bought both a and b
  - the number of peoples who bought a
  - the number of peoples who bought b

- It is easy to count the number of peoples who bought a or b. Because we already have the item list.
- We need to count the number of peoples who bought both a and b.

※ Cosine similarity :

$\cos(a,b) = a \cdot b / (|a| \cdot |b|)$

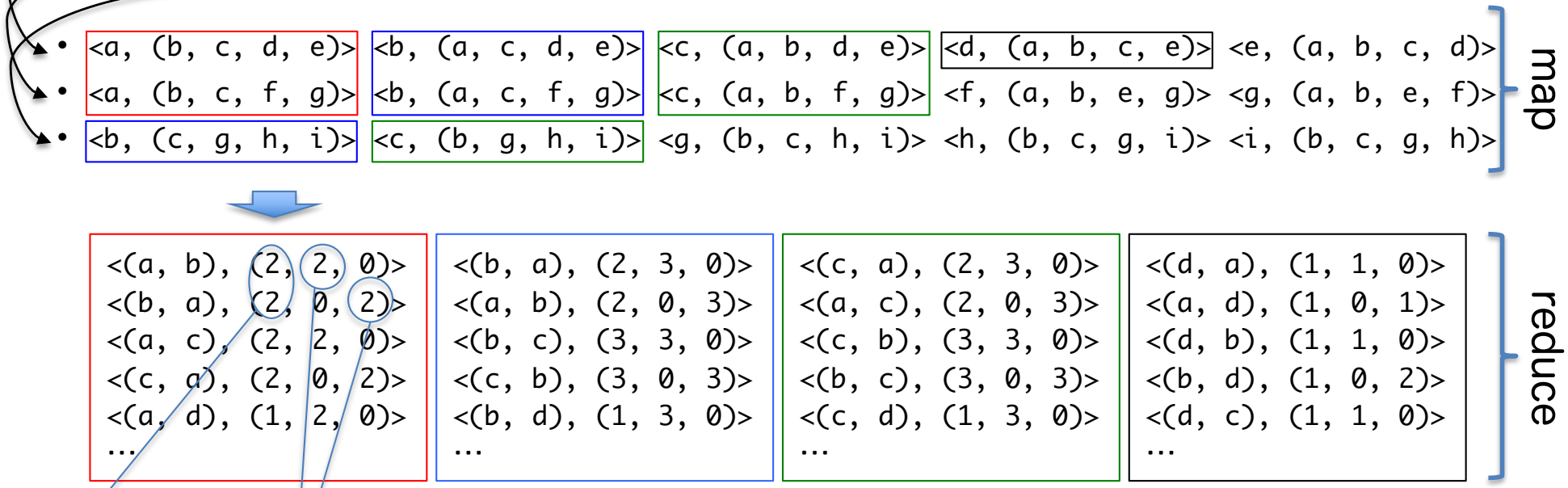
= num of users who access both / num of users who access a or b

協調フィルタリングを考える (7)
Collaborative Filtering (7)

- 両方を購入した人数を計算するためには、次のようにすれば良い。
  To count num of peoples who buy both…
  - ユーザが買ったもののリストを作成する。/ Make a list peoples buy.
    - α (a, b, c, d, e)
    - β (a, b, c, f, g)
    - γ (b, c, g, h, i)

α, β, γ: People
a, b, c, … : item

  - その上で、ある品を買った人が他に何を買っているのかを調べる。



協調フィルタリングを考える (8)
Collaborative Filtering (8)

– 品物のペア毎に集計し、類似度を計算する。
Totaling by item pair. Make similarity.

<(a, b), (2, 2, 0)>	<(b, a), (2, 3, 0)>	<(c, a), (2, 3, 0)>	<(d, a), (1, 1, 0)>
<(b, a), (2, 0, 2)>	<(a, b), (2, 0, 3)>	<(a, c), (2, 0, 3)>	<(a, d), (1, 0, 1)>
<(a, c), (2, 2, 0)>	<(b, c), (3, 3, 0)>	<(c, b), (3, 3, 0)>	<(d, b), (1, 1, 0)>
<(c, a), (2, 0, 2)>	<(c, b), (3, 0, 3)>	<(b, c), (3, 0, 3)>	<(b, d), (1, 0, 1)>
<(a, d), (1, 2, 0)>	<(b, d), (1, 3, 0)>	<(c, d), (1, 3, 0)>	<(d, c), (1, 1, 0)>
…	…	…	…

map

- <(a,b), (2,2,0)> <(a,b), (2,0,3)> → <(a,b), (2,2,3)> → 0.816
- <(b,a), (2,0,2)> <(b,a), (2,3,0)> → <(b,a), (2,3,2)> → 0.816
- …
- <(b,c), (3,3,0)> <(c,b), (3,3,0)> → <(b,c), (3,3,3)> → 1.000
- …
- <(b,d), (1,3,0)> <(b,d), (1,0,1)> → <(b,d), (1,3,1)> → 0.577
- …

reduce

課題2
Assignment 2

- Wikipediaの統計を解析し、レポートを作成しなさい。
- Analyze Widipedia statistics and make a report

① 解析の目的について述べなさい。/ What is your purpose of analysis?
※できるだけ授業でやったこと以外をやってみてください。

② 解析の手順について述べなさい。/ Process of Analysis

③ わかったことを述べなさい。/ What do you find in the analysis?

④ やってみた感想を述べなさい。/ What kind of impression did you have?

- 注意 / Notice
  - 必ず、氏名・学籍番号を記載のこと。/ You MUST put your name and student id number.
  - いわゆるレポートの形で作成すること。(スライド不可) / Slide is not acceptable. Make a report.

- 締め切り: 6/8（金）23:59 / Deadline: 8th Jun (Fri) 23:59